REPORT RESUMES

ED 012 829 24 AA 000 189

THE EFFECTIVENESS OF COLLEGE-LEVEL INSTRUCTION IN FRESHMAN COMPOSITION.

BY- JEWELL, ROSS M. AND OTHERS
STATE COLLEGE OF IOWA, CEDAR FALLS
REPORT NUMBER BR-5-0803 PUB DATE DEC 66
CONTRACT OEC-SAE-4-10-053
EDRS PRICE MF-$0.50 HC-$3.64 91P.

DESCRIPTORS- *COMPOSITION (LITERARY), *COLLEGE STUDENTS, *WRITING, *PERFORMANCE, *EFFECTIVE TEACHING,

THE WRITING PERFORMANCE OF STUDENTS COMPLETING FRESHMAN COMPOSITION WAS COMPARED WITH THE WRITING OF STUDENTS NOT TAKING FRESHMAN COMPOSITION WHEN BOTH HAD BEEN IN COLLEGE THE SAME LENGTH OF TIME. FOR THE INVESTIGATION, 325 STUDENTS TAKING COMPOSITION WERE MATCHED WITH STUDENTS NOT TAKING COMPOSITION ON THE BASIS OF AGE, SEX, SCORES ON A WRITTEN THEME, ON THE "COLLEGE ENTRANCE EXAMINATION BOARD ENGLISH TEST," AND ON THE "COOPERATIVE ENGLISH TESTS--ENGLISH COMPOSITION." STUDENTS WERE TESTED AT THE START, AT THE END OF THE FIRST SEMESTER, AT THE END OF THE SECOND SEMESTER, AND AT THE END OF THE FOURTH SEMESTER. RESULTS SUSTAINED THE HYPOTHESIS THAT THE WRITING PERFORMANCES OF STUDENTS WHO COMPLETE A YEAR OF COMPOSITION DO NOT DIFFER SIGNIFICANTLY FROM THAT OF STUDENTS WHO HAVE HAD NO COMPOSITION. THE INVESTIGATORS PLANNED TO CONDUCT A SECOND PHASE OF THIS PROJECT TO STUDY THE COMPOSITION WRITING SKILLS OF STUDENTS AT FIVE OTHER INSTITUTIONS. (TC)

ED012829

AA000189

5-0803

INTERIM REPORT
Project 2188, Amended
Contract SAE-OE-4-10-053 Amended

THE EFFECTIVENESS OF COLLEGE-LEVEL INSTRUCTION IN FRESHMAN COMPOSITION

December 1966

U.S. DEPARTMENT OF
HEALTH, EDUCATION, AND WELFARE

Office of Education
Bureau of Research

THE EFFECTIVENESS OF COLLEGE-LEVEL INSTRUCTION IN FRESHMAN COMPOSITION

Project 2188, Amended
Contract SAE-OE-4-10-053 Amended

Ross M. Jewell, Director
John Cowley, Gordon Rhum

December 1966

The research reported herein was performed pursuant to a contract with the Office of Education, U. S. Department of Health, Education, and Welfare. Contractors undertaking such projects under Government sponsorship are encouraged to express freely their professional judgment in the conduct of the project. Points of view or opinions stated do not, therefore, necessarily represent official Office of Education position or policy.

State College of Iowa

Cedar Falls, Iowa

CONTENTS

Page

Purposes and Procedures 1
  Statement of the Problem 1
  Pilot Phase 2
  Related Research 2
  Procedure 15
    Evaluative Instruments 16
    Establishing Matched Pairs 16
    Theme Evaluation 18

Results 20
  September, 1963, Test Performance of Original and Persisting Subgroups 20
  Criterion Scores--September, 1963, through May, 1965 . 22
  Comparison of Criterion Scores--Sample Available January, 1964 24
  Comparison of Criterion Scores--Sample Available May, 1964 27
  Comparison of Criterion Scores--Sample Available May, 1965 35
Intercorrelations Among Variables 41
  September, 1963--Total Group 41
  January, 1964--End of First Semester 45
  May, 1964--End of First Year 48
  May, 1965--End of Second Year 52

Reliability of Criterion Measures 52
  Cooperative English Tests: English Expression 52
  The College Entrance Examination Board English Composition Test 54
  The Theme 54

Performance by Sex and Ability Level 56
  Performance by Sex 56
  Performance by Ability Level 59

Conclusions and Observations 63
  The Basic Findings 63
  Performance by Sex and by Ability Level 65
  Observations 65

Summary 67

References 69

Page

Appendices A-1
  Appendix A--Theme Topics and Instructions A-1
  Appendix B--Choice of Experimental Design B-1
  Appendix C--Procedure for Evaluating Themes C-1

LIST OF TABLES

Table — Page

I. Achievement as of September, 1963, of the Entire Entering 1963-64 Freshman Class and of Various Persisting Subgroups of That Class 21

II. Performance of Available Matched Pairs of Students on Three Criterion Measures at Beginning, Middle, and End of First Year and End of Second Year of College 23

III. The Performance of 166 Matched Pairs of Students on the *Cooperative English Tests: English Expression* at the Beginning and at the End of the First Semester of College; Differences in Means, and t-Ratios 25

IV. Performance of 166 Matched Pairs of Students on the *College Entrance Examination Board English Composition Test* at the Beginning and at the end of the First Semester of College; Differences in Means, and t-Ratios 26

V. Performance of Approximately 166 Matched Pairs of Students on the Sum of Two Theme Ratings at the End of the First Semester of College; Difference in Means, and t-Ratio 28

VI. The Performance of 113 Matched Pairs of Students on the *Cooperative English Tests: English Expression* at the Beginning of the First Semester, at the End of the First Semester, and at the End of the Second Semester; Differences in Means, and t-Ratios 29

VII. The Performance of 113 Matched Pairs of Students on the *College Entrance Examination Board English Composition Test* at the Beginning of the First Semester, at the End of the First Semester, and at the End of the Second Semester; Differences in Means, and t-Ratios 32

VIII. The Performance of 110 Matched Pairs of Students on the Total of Two Theme Ratings at the End of the Second Semester; Differences in Means, and t-Ratios 34

Table Page

IX. The Performance of 31 Matched Pairs of Students on the *Cooperative English Tests: English Expression* at the Beginning of the First Semester, at the End of the First Semester, at the End of the Second Semester, and at the End of the Fourth Semester; Differences in Means, and t-Ratios 36

X. The Performance of 31 Matched Pairs of Students on the *College Entrance Examination Board English Composition Test* at the Beginning of the First Semester, at the End of the First Semester, at the End of the Second Semester, and at the End of the Fourth Semester; Differences in Means, and t-Ratios 38

XI. The Performance of 31 Matched Pairs of Students on the Total of Two Theme Ratings at the End of the Fourth Semester (May, 1965); Differences in Means, and t-Ratios 40

XII. A Summary Comparison of Test Score Means of Experimental Subgroup and Control Subgroup at Three Testing Points: Indication of Statistically Significant Differences for Specified Subsamples 42

XIII. Intercorrelations Among 9 Variables for 910 New Freshmen Entering the State College of Iowa, Fall Semester, 1963-64 43

XIV. Intercorrelations Among 15 Variables for 332 Students (166 Matched Pairs) at the End of the Fall Semester, 1963-64 46

XV. Intercorrelations Among 15 Variables for 166 Experimental Students at the End of the Fall Semester, 1963-64 47

XVI. Intercorrelations Among 16 Variables for 166 Control Students at the End of the Fall Semester, 1963-64 49

XVII. Correlation Between Selected Pairs of Variables for the 113 Matched Pairs of Students Completing the Spring Semester, May, 1964 51

Table Page

XVIII. Correlation Between Selected Pairs of Variables for the 31 Matched Pairs of Students Completing the Spring Semester, May, 1965 53

XIX. Frequency Distribution of the Difference in Two Independent Ratings Assigned to Each of 1,070 Themes 55

XX. Performance of 113 Matched Pairs of Students, by Sex, on Three Criterion Measures at the Beginning, Middle, and End of the First Year of College 58

XXI. Gains on *College Entrance Examination Board English Composition Test* During the College Freshman Year at Each of Four ACT-English Ability Levels 61

FOREWORD

The study presented here is the complete report on the pilot phase of Research Project 2188, amended, under a contract between the United States Office of Education and the State College of Iowa. The data for the report on the second phase have been gathered and the report itself will be forthcoming in 1967.

Though more complete acknowledgements of assistance may be made in the final report, it is not too early for the investigators to express their appreciation to the many individuals and organizations which have assisted in bringing it to fruition. First among these must be Dr. J. W. Maucker, President of the State College of Iowa; Dean William C. Lang; Dr. H. W. Reninger, Head of the Department of English Language and Literature; Dr. Marshall Beard, Registrar; and Mr. Paul Mahon, until recently in charge of Data Processing. Had the administration of the college not had the courage to allow students to omit a course frequently considered to be vital to their success in college and in life, this project could not even have begun.

We have also enjoyed the cooperation of the College Entrance Examination Board, which made available the *CEEB English Composition Test*, one of our three test instruments.

Last, we owe a great debt of gratitude to the students who participated in the investigation. They were at all times cooperative and helpful, whether they were in the experimental or the control group. We hope that the impact of the findings here reported will justify their cooperation.

PURPOSES AND PROCEDURES

Statement of the Problem

Research in college composition has not been plentiful, and most of the studies reported have concentrated on comparing some innovation with a standard procedure. Variables have ranged from the number of papers written through the amount of teacher comment on each paper to the influence of such subjects as rhetoric and grammar on the performance of the student. In every case the other element in the comparison was the particular arrangement of freshman composition at the institution in which the research was done. Seldom has a statistically significant difference appeared, and the difficulty is that, even where it has, the difference has been between a particular innovation and what might be termed standard procedure. A tacit assumption in all such research has been that the "standard" course improved student writing and the question was whether the innovation would produce a result different from that produced by the standard course. These investigations seldom included comparisons of the results with an arrangement involving no formal instruction in English composition.

A second difficulty with the research reported has been that the statistical comparisons involved a relatively small number of students. The question is always present as to whether the sample employed is sufficiently large and broadly based to be reasonably representative of a given group--for example, all entering college freshmen in a substantial number of American colleges. In those few instances in which a statistically significant difference has been found, the degree to which generalizations beyond the samples investigated may be made is uncertain.

The present investigators decided to attempt to overcome both of these deficiencies. They planned to compare students who had received no instruction of the sort generally given in freshman composition with comparable students who had received such instruction. In order to develop statistics for a reasonably broad and a reasonably diverse population, they planned to engage several institutions in replicating the experiment. This procedure would give a numerical, geographical, and academic variety to the population. If the results at all participating institutions were in agreement, the conclusions could be stated with considerable force. If the results among the institutions varied, directions for future investigation might be indicated.

The goals of the investigation, then, were to test two hypotheses:

(1) That the writing performance of the students enrolled in a freshman composition sequence is not significantly different from the writing performance of students not enrolled in a freshman composition sequence when the two groups have been in college for an equal length of time.

(2) That the results obtained in (1) will be present in other colleges or universities.

A by-product of the testing of the hypotheses would be the accumulation of statistics based upon a reasonably large and diverse sample of students who had received no instruction in college freshman composition. Such a set of statistics might prove useful in providing a realistic and stable base for investigating the effect of innovation as well as of the "standard" course itself. Meaningful use of these statistics could be made only if the investigators testing an innovation utilized the evaluative instruments employed in the present investigation.

Pilot Phase

The present report covers the pilot phase of a two-phase project. It is based upon experiences at the State College of Iowa from September, 1963, through May, 1965. The second phase of this project will involve the performance of students at five institutions: the University of Colorado, the University of Iowa, Kent State University, Northern Illinois University, and the State College of Iowa, from September, 1964, to May, 1966.

Related Research

No research has come to the investigators' attention which is directly comparable to the present study. Nearly all the research compares some innovation with a standard procedure. Such studies ordinarily vary the frequency of writing in the composition course as the experimental variable. Most of these obtained no statistically significant differences in the performance of the groups of students at the end of instruction. A summary of projects with some relevance to the current study is given below.

Arnold, Lois. Effects of Frequency of Writing and Intensity of Teacher Evaluation upon Performance in Written Composition of Tenth Grade Students (Cooperative Research Project Number 1523), Tallahassee: Florida State University, 1963, University Microfilms No. 63-6344.

Miss Arnold conducted her research in 1961-1962 at two Florida high schools, in each of which a teacher was scheduled to teach four groups of students in the tenth grade. The four groups at each school were average classes, determined by sectioning on the basis of scores on the following tests: *Pintner General Ability Test*, *Metropolitan Achievement Battery*, *School and College Ability Test*, and *Differential Aptitude Tests*. Students were classified as low average, middle average, or high average on the basis of the *DAT* scores. Nothing is said of student-to-student matching. The experiment lasted for the school year. Each teacher at each school used four teaching methods, a different one for each of her four classes as follows:

1. Infrequent writing, moderate evaluation: one theme, approximately 250 words, each six weeks. Evaluation was concentrated on one matter each time: once on sentence structure, once on organization, etc.
2. Frequent writing, moderate evaluation: some writing four times a week, varying from two sentences to two pages or more. The evaluation was handled as in 1 above.
3. Infrequent writing, intensive evaluation: one theme each six weeks, approximately 250 words. Every error in usage, sentence structure, and mechanics was marked and detailed comments written on the paper. Students corrected all errors, revised or rewrote until the paper was satisfactory.
4. Frequent writing, intensive evaluation: one 250-word theme weekly, evaluated meticulously as in 3 above (pp. 40-2).

Two evaluative instruments were used, *STEP Essay Tests* and *STEP Writing Tests*, the former a writing test, the latter an objective test. Both were administered at the beginning and at the end. Three experienced (former) English teachers independently rated the *STEP Essay Tests*, the pretests in December and January, and the post-tests in May and June.

Miss Arnold reached four conclusions:

1. There is no assurance that intensive evaluation is any more effective than moderate evaluation in improving the quality of written composition.
2. It must not be assumed that frequent practice is in itself a means of improving writing.

3. There is no evidence that any one combination of frequency of writing and intensity of evaluation is more effective than another.

4. There is no indication that frequent writing and intensive evaluation are any more effective for one ability level than are infrequent writing and moderate evaluation (p. 62).

In this study there was no significant difference between the sexes.

The SCI investigators wonder whether graders might have evaluated more alike had they conferred on an occasional paper (Four correlations were in the .50's, the others being .62 and .76), and why, in a gains study, all themes were not scored at a single time with prethemes and post-themes mixed. A table showing comparisons of the terminal data only would also have been helpful. That is, how did the groups compare at the end regardless of gains?

Buxton, Earl W. "An Experiment to Test the Effects of Writing Frequency and Guided Practice upon Student's Skill in Written Expression." Unpublished Ph. D. dissertation, Stanford University, 1958. University Microfilms 58-3596. [As reported in Braddock, et al. Research in Written Composition. Champaign, Illinois: NCTE, 1963, pp. 58-70.]

This experiment involved 257 students in the University of Alberta who were enrolled in a special "one-year 'emergency' course designed to train teachers for Alberta schools." All 257, who constituted the entire enrollment in the emergency program, carried the same courses (a "canned" schedule). The total group was divided into six classes: two control classes, in which students did no extra, out-of-class writing; two writing classes, in which students wrote a 500-word paper each week as an extra out-of-class assignment for a total of sixteen weeks; two revision classes, in which students did the same amount of writing on the same assignments as the writing classes. Writing classes were not required to write on the assigned topic and received only a brief paragraph of teacher comment at the end of each theme; there was no marking of errors nor commenting in the margin, and students were not asked to do anything with the papers after getting them back. The revision classes were required to write on the assigned topic and papers were marked in terms of unity, organization, logic, correctness, and such matters, with a general comment at the end. Students in the revision classes were asked to correct and revise their papers in class on the day the papers were returned and discussed. The teacher was present to give aid.

Criterion measures were two parts of an earlier edition of the Cooperative English Tests: "Mechanics of Expression" and "Effectiveness of Expression" (alternate forms before and after), and a theme. Each of two readers assigned a "content" score and an "error" score to each theme. The content score was based on fifteen factors with some factors weighted more than others. A maximum potential score was allotted for each factor. Each reader determined how much of that maximum to assign to that factor in each paper. The error score was determined by counting errors in spelling, punctuation, or mechanics. The points assigned for each of the fifteen factors in a paper by each reader were added; then the count for errors was subtracted from that. The scores for the two readers were averaged, and that mean was arbitrarily divided by three to get a usable scaled score.

The results of Buxton's study show that the revision students--those whose papers were carefully marked and who were required to revise them--made a significantly greater gain in writing achievement as measured by the themes during the seven months of the study than did the writing students--those who wrote the papers but did not revise them. There was a more significant difference in gain scores between the revision students and the control students, who wrote none of the themes; this difference favored the revision students. Concomitant conclusions: theme ratings are reliable if the raters are thoroughly practiced in their system and frequently check on what they are doing, and (since there was no significant difference between the groups on the objective test scores) the theme ratings in this study measure something that the particular objective test used did not measure.

It is not clear whether the division into groups took into account the balance of men and women. If, for example, the revision classes had more women than either of the other two groups, that could affect the results.

Heys, Frank, Jr. "The Theme-a-Week Assumption: a Report of an Experiment," English Journal, 51 (May 1962), 320-22.

This experiment dealt with varying the amount of writing and the amount of reading in high school English classes. Two classes in each of the four high school grades were "as closely matched as was possible under the normal sectioning practices of the school." The two classes in each grade were taught by the same teacher; one was designated as the writing class and the other as the reading class. Students in each writing class wrote a theme a week. After it was closely graded, the students corrected or rewrote it. Students in each reading class wrote a theme every three weeks, and spent one class day a week reading books of their own choice. Nothing is said concerning grading or rewriting of the reading-class papers. Evaluation instruments consisted of the STEP writing test and a theme, one of each administered at the beginning and at the end of the experiment. The themes were evaluated by three ETS readers using a nine-point scale.

The students in reading classes achieved a slightly greater improvement in writing scores than did those in writing classes. Generalizations arrived at by the investigator:

1. Frequent writing practice probably yields greater dividends in grade 12 than in grades 9, 10, 11.

2. Frequent writing practice probably yields greater dividends with low groups than with middle or high groups.

3. Frequent writing practice with low groups probably yields greater dividends within the area of content and organization than within the area of mechanics or of diction and rhetoric.

4. The claim that "the way to learn to write is to write" is not substantiated by this experiment.

5. The claim that ability to write well is related to the amount of writing done is not substantiated by this experiment.

6. For many students reading is a positive influence on writing ability.

7. The influence of reading on the ability to write appears to be a separate factor, not directly related to the teacher's personality and enthusiasm (p. 322).

It is not clear how the fourth generalization is supported by the experiment. Since all students in the experiment wrote themes, how can it be inferred that the data failed to support the notion that students learn to write by writing? Furthermore, Heys does not indicate whether the improvement mentioned was statistically significant.

Kincaid, Gerald L. "Some factors Affecting Variations in the Quality of Students' Writing." Unpublished Ed.D. dissertation (Michigan State University, 1953). University Microfilms No. 5922.

This experiment attempted "to determine whether a *single* paper written on a given topic at a particular time [italics Kincaid's] can be considered as a representative sample of his [the student's] writing ability--and thus provide a valid basis for evaluating ability at any time in a writing course." It is of interest, not because it deals with a directly related problem, but because it has implications for any study using theme readers to evaluate results. A group of 80 college students was divided into four subgroups, each of which wrote two papers in one two-hour session on the same day and another two papers in a similar session a week later. Three topics were used: Groups A and C wrote on topics 1 and 2 each time (both argumentative); groups B and D wrote on topics 1 and 3 each time (one argumentative, one expository). Groups A and B wrote each time without examination pressure (papers not counted toward grade); groups C and D wrote without pressure once, and with it the other time (papers counted on term grade the first time and not counted on term grade the second time). Papers were rated by three instructors selected from the freshman staff, the rating being made on a ten-point scale (1 unsatisfactory, 10 superior) on each of five categories: grammatical conventions, sentence structure, diction, organization, and content. The score for a paper could lie between 10 and 50; it was determined by computing the mean of the two closest ratings; if the two extreme ratings were equidistant from the middle rating or if the two closest ratings were more than five points apart, the mean of all three was used.

Kincaid drew the following conclusions from this study:

> 1. . . . the findings from this study cast considerable doubt upon the justification of the customary practice of using five letter-grades to designate [individual] achievement in a writing course when a single paper provides the basis for that designation (p. 97).
>
> 2. If an evaluation of over-all or average improvement is all that is desired, it can be obtained from a single sample of each student's writing for a pre-test and a post-test . . . (p. 99).
>
> 3. . . . in order to develop a program for evaluating individual student improvement in writing (for strong as well as for weak students), it would be advisable to obtain

> several samples of writing by each student--samples of writing on different topics on the same day and on the same topics on different days. And such samples should be obtained for both the pre-test and the post-test (p. 99).

Two matters impress the present investigators: 1) The theme topics used by Kincaid were simpler than those used in the State College of Iowa investigation. If more difficult topics had been used by Kincaid the results might have been different. 2) The findings of the Kincaid investigation support the use of group average scores on a single pretheme and a single post-theme.

Kreisman, Arthur, *et al*. *Pilot Study in English*. Mimeographed report and dittoed summary of statistics. Ashland, Oregon: Southern Oregon College, 1963 (no pagination).

This is the report of a pilot study designed "to investigate techniques and writing skills as a possible means of establishing the basis for a more extensive research program." It is interesting because the results led the Oregon investigators to abandon further experimentation, and because one of those investigators suggested a study like the State College of Iowa study. In the Oregon study, both college freshmen and high school students were involved. Control and experimental groups were matched at both levels: the 89 college students on the Verbal and Quantitative scores on SAT, the total score on SCAT, and the sum of two ratings on the STEP Essay Test; the 108 high school students on the score on the California Test of Mental Maturity and the sum of two ratings on the STEP Essay Test. Both control and experimental students were in each class. The amount of writing actually done is not clear. In one place Kreisman says that the experimental students wrote three themes, the control students nine themes. He then says that the college-student experiment lasted for one term, the high school experiment lasted for the year. He says further that the experimentals wrote once a month, the controls once a week. Evaluation was based upon comparison of the STEP essay ratings at the beginning and at the end of the experiment.

There was no significant difference between the college experimental and control groups. The results for the high school groups varied. There was a significant improvement for the below-average high school students in the control group (more writing); there was a slight (non-significant) drop in achievement for the above-average students in the control group (more writing). There was no significant difference in the experimental group (less writing). Dr. Cloer, the statistician, wrote: "It would appear that the principal beneficiaries of the experience in writing were those subjects of below-average ability or those who might be called 'under-achievers,' . . ."

Comments quoted from Kreisman:

1. No adequate instrument for testing [composition] seems available.
2. The difficulty of obtaining a sufficient number of students to make the experiment valid was one of the major obstacles.
3. . . . a purely quantitative experiment has little chance of being valid.
4. . . . one term of writing practice is not sufficient to form a foundation for judgment regarding the development of writing ability.
5. . . . frequency may indeed be a factor in the development of writing ability.
6. . . . all experiments of this nature are of no value and invalid on an a priori basis.

In the light of the State College of Iowa study, the following additional comments are of special interest, the first by Kreisman, the second by Cloer, the statistician: "The emphasis that we thought might be fruitful [for future research] would be one which dealt with student-teacher relationships or with maturation of students regardless of the courses they took," and "Perhaps a better 'experimental group' would be one that did no writing (in English classes) over the experimental period."

McColly, William and Robert Remstad. Comparative Effectiveness of Composition Skills Learning Activities in the Secondary School (Cooperative Research Project 1528). Madison: University of Wisconsin, 1963.

This study attempts to answer three questions:

Does more writing alone result in better writing?

Do more of "functional non-writing composition learning activities" (practical instruction: working with

> student-written papers, emphasizing spelling, proofreading, revision, etc.; group discussion; teacher evaluation and comment) result in better writing?
>
> Does tutoring with immediate feedback (having the teacher present while the writing is being done and advising the student during the process) result in better writing? (p. 18)

To answer the first question, dealing with the effect of the quantity of writing on improvement in writing, the investigators used two classes in the eighth grade and two classes in the ninth grade. To answer the questions relating to "functional non-writing activities" and immediate feedback (tutoring), three classes in each of the tenth, eleventh, and twelfth grades were used. Covariance techniques and, to the extent possible, random selection of samples were employed.

To explore the effect of the amount of writing on improvement in writing, *control* classes in the eighth and ninth grades wrote a theme a month; *experimental* classes wrote a theme a week. All other class activities and assignments were the same. During the year, the eighth-grade *control* classes wrote 9 themes and the eighth-grade *experimentals* wrote 35 themes. The ninth-grade *control* classes wrote 8 themes, the *experimentals*, 34.

To study the effect of non-writing activities and tutoring, one *control* class (a monthly theme with functional instruction), and two *experimental* classes (weekly theme and functional instruction), were organized at each grade level. About 9 writing tasks with functional activities were completed in the *control* classes, about 34 in the *experimental* classes. There were no individual conferences or "tutoring" activities in the first of these *experimental* classes in each grade. There were about 27 regular "tutoring" sessions in the second *experimental* class in each grade. Thus, a ratio of 4-1 was maintained in writing tasks with functional activities between the *experimental* and *control* classes.

Criterion and covariate measures for all students in the experiment included: SCAT (IA, IIA, IIIA), Nelson-Denny Reading, ITED ("Correctness and Appropriateness of Expression" and "Ability to Interpret Literature"), previous English GPA, overall GPA, and writing samples, two written before the experiment and two written at the end.

Based on this experiment, the answer to the first question is *no*. Results indicated that increase in the amount of writing by itself has no significant effect upon the writing proficiency of high school students. Again, based on this experiment, the answer to the second question is affirmative; the answer to the third question is negative. Experimental classes with weekly theme and functional instruction improved significantly compared to the control classes. The experimental classes with tutoring scored, at the end of the experiment, about half way between the control classes and experimental classes without tutoring.

Rohman, D. Gordon and Albert Wlecke. Pre-writing: The Construction and Application of Models for Concept Formation in Writing (Cooperative Research Project No. 2174), East Lansing, Michigan: Michigan State University, 1964.

This is one of the very few studies that have resulted in a statistically significant difference between control and experimental groups. Six sections of a college sophomore course in expository writing with an emphasis on pre-writing activities constituted the experimental group. Three sections were taught each quarter for two quarters. The rest of the students enrolled in the same course (11 sections in the Winter term, 10 in the Spring term), constituted the control group. The total number of students involved in the experiment is not disclosed. The experimental course contained six units: 1. The role of the writer. 2. The escape from category (the concrete rather than the abstract). 3. The escape from cliché (avoiding someone else's way or words). 4. Dynamic relationship to the subject (an urgency to express what the writer has "discovered"). 5. Concrete analogy (expressing one's "discovery" by comparison with something like it). 6. Refinement (finishing the essay). Three major techniques were used: keeping a journal, meditation, and use of analogy. The control sections were taught as each teacher wished to teach them, with the exception that all instructors of the control sections assigned two 500-word themes on topics used in the experimental sections. These themes were used in the evaluation.

Evaluation of the experiment involved four devices: 1. statements written by students in answer to the question: What did you like or dislike about the course?, 2. statements by the teachers who taught the course, 3. "objective" evaluation by readers who did not teach the course, and 4. "subjective" evaluation by teachers who did not teach the course. No objective testing was reported.

Evaluation by students was strongly favorable. Major items were that the course was enjoyed, that it developed freedom in writing and in the discipline of writing and thinking, that criticism of student writing led to involvement in the process of writing, that attitudes toward writing had changed (regarding, for instance, the relationship between thinking and writing), that the use of analogy led to greater concreteness and clarity. Negative criticisms, which were relatively few, included the following: the course was too short; it was too piecemeal; not enough grades were given; class criticism was too negative; the journal was an invasion of privacy; the use of analogy was mechanical.

Instructors gave a number of reactions to the experiment, but their enthusiasm tended to center on three matters: the journal as a device to stimulate students to meditate about their experiences as well as to formulate their meditations in writing, the emphasis on the pre-writing process, and the freshness and soundness of the writing done.

The essays for "objective" evaluation were selected from the total submitted by *control* and *experimental* subgroups on the two topics used by both subgroups. There were 226 *experimental* and 409 *control* essays evaluated. No information is given concerning how these essays were selected. Essays were judged on a four-point scale: 4. superior, 3. above average, 2. below average, 1. incompetent. Three standards, unity, coherence, and emphasis, were guides for the readers. There were 11 readers, four high school teachers and seven college teachers. They worked in teams of 8, three who read at the first session not reading at the second, and three others substituting for them at the second. Each theme was read twice. About 85% of the grades assigned were either the same for each theme or only one point different, indicating that the grading was relatively reliable. The results showed a statistically significant difference between the *experimental* and *control* groups in favor of the *experimentals*.

Four members of the English staff not involved in the experiment read the papers "subjectively." They were given a randomly selected sample of 50 *experimental* and 50 *control* themes. Rohman and Wlecke informed these readers concerning which set was *experimental* and which was *control*. Some investigators would not have done that. The readers were asked to answer a series of three questions: "Which set of essays seems to have more originality and in what ways? Generally, in which set of essays does it seem more important for the writers to express themselves and not be misunderstood? Which set of essays gives the greater sense of form?" (pp. 130-1) In addition, the readers were asked a series of specific questions concerning only the *experimental* essays, such as: "Do the techniques employed in the *experimental* essays--the meditation in the 'Loneliness' essays, and the analogy in the 'Coming of Age' essays--seem to provide a more coherent means for the instructor to gauge the success or failure of an essay?" All four readers gave the experimental group of essays the higher rating.

Rohman and Wlecke leave so many questions unanswered that the report is difficult to interpret. How many students were in each sample? Were the students of the experimental sections similar in ability to those in the control sections? Did either sample have appreciably more women than the other? How were the themes that were evaluated selected? Do the 226 experimental themes represent a sampling comparable to the 409 control? Would a sampling of the control students have written as enthusiastically of their course as the experimentals did? To what degree did the Hawthorne effect operate? What implications has this study for composition programs generally?

Sutton, Joseph T. and Eliot Allen. *The Effect of Practice and Evaluation on Improvement in Written Composition.* (Cooperative Research Project No. 1993). Deland, Florida: Stetson University, 1964.

This study randomly divided college freshmen into five groups. The first two of these (Groups I and II) served as controls. During the period of the experiment, these two groups received no instruction in composition and wrote no papers except the six criterion themes which provided the "before" performance and the six criterion themes which provided the "after" performance. Group I wrote all twelve themes within a four-week period at the beginning of the semester. Group II wrote the first six criterion themes the first two weeks of the semester and the second six criterion themes the last two weeks of the semester. Groups III through V were the experimental groups, and all wrote six criterion themes the first two weeks and another six the last two weeks (as did Group II). In the ten-week interval between the writing of criterion themes, Group III wrote no papers but did evaluate four peer papers each week; Group IV wrote one theme each week which was evaluated by the members of Group III; and Group V wrote one class theme each week which was evaluated by a "professor."

Five readers read each theme twice, once to rate it, once to rank it in an order of excellence relative to the other eleven themes by each writer. Rankings were based on five criteria: ideas, mechanics, wording, form, and flavor, each one of which was scored on a five-point scale. A total for the six "before" themes for each student as graded by all five graders, divided by thirty (6 themes x 5 graders) gave an average score for each writer. The same was done for the six "after" themes, and the averages were compared.

Particularly in relation to the State College of Iowa study, Sutton and Allen's enterprise is interesting. First, none of the students in any of the groups received direct instruction in composition. Such instruction as Groups IV and V received came from the marks and comments on their papers. Group III gained experience in editing, though uninstructed in the procedure. Groups I and II had no experience whatsoever with composition except the twelve criterion themes. Thus, to a degree this study is similar to the present one in that no direct instruction in freshman composition was given and that some of the groups wrote only the criterion themes. It is different from the present study in that there was not a direct comparison between those completing a freshman program of writing instruction and others not in the freshman English course at all.

The results in the Sutton and Allen study showed an unusual inconsistency between the themes and the objective tests. In theme performance, the members of the five groups showed a significant decline during the experimental period. A decline was observed for the five groups combined and for each group separately. This decline was, of course, unexpected. The authors, in speculating about its source, state: "Unfortunately, it appears that the very procedure necessary to secure such stability [among the theme performances] introduced other factors that may have had a deleterious influence on the results." The frequency of writing of test themes which were neither returned to the student nor commented on seems, in the opinion of Sutton and Allen, to have created an attitude of boredom and impatience among the students. On each of the two objective tests, the *Cooperative English Tests: English Expression* and the *College Entrance Examination Board English Test*, the students showed significant improvement. This was true for the five groups combined, and there was no significant variation among the five groups in this respect.

Wolf, Melvin H. *Effect of Writing Frequency upon Proficiency in a College Freshman English Course*. (Cooperative Research Project 2846), Amherst, Massachusetts: University of Massachusetts, 1966.

This study involved six "regular" sections of college freshman composition and four remedial sections. Two of the regular sections, designated *experimental-high frequency*, wrote 39 themes in the school year; two sections, designated *experimental-low frequency* wrote 8 themes in the year; two sections, designated *control*, wrote 15 themes in the year, the usual number in freshman composition at the University of Massachusetts. Two remedial sections, designated *experimental-high frequency*, wrote 20 themes in one semester; the other two, designated *control*, wrote 8 themes in one semester. These themes were carefully evaluated by the instructors and were revised and resubmitted by the students. The objective test used was *Cooperative English Tests, Form 1C*. Six themes were used as tests: two written at the start, two at the end of the first semester, and two at the end of the second semester. The remedial students, being in the study only one semester, wrote only the first four test themes. Evaluation of the test themes was done by 10 instructors under the direction of an experienced instructor who had been a reader for the Educational Testing Service. Wolf drew two conclusions: 1) writing proficiency did not improve with the increase in frequency of writing, 2) there was a high correlation between the scores on objective tests of grammar and mechanics and scores of themes as determined by the reading team. Since COOP has a section on mechanics and a section on effectiveness but usually yields a single score, it is not clear how the second conclusion was arrived at.

Procedure

The overall design of the pilot project involved selecting experimental and control groups at the State College of Iowa and testing them on four different occasions: the beginning of the freshman year (September, 1963), the end of the first semester (January, 1964), the end of the first year (May, 1964), and the end of the second year (May, 1965). Members of the experimental group received no instruction in freshman composition; members of the control group did receive instruction in freshman composition. The performance of these groups was compared at each testing period to determine whether the differences in their performance on the criterion measures were significant. Care was taken that the members of each group would be representative of the total freshman class entering the State College of Iowa in September, 1963. Members of both experimental and control groups pursued a normal academic program except that the experimentals omitted the freshman composition course. The experimental group substituted other general education courses for freshman composition and thus took some of those courses a semester earlier than the control group did, or carried a course in their majors earlier than most of the control students did.

Evaluative Instruments

Three tests of performance in composition were used: the Cooperative English Tests: English Expression (COOP), the College Entrance Examination Board English Composition Test (CEEB), and a theme. The first two are objective tests. The COOP appealed to the investigators because it had been employed in previous research at the State College of Iowa and seemed to serve as a reasonably satisfactory indirect measure of student writing ability. The CEEB, unlike the COOP, is a "secure" test. It is changed from administration to administration and a serious attempt is made to assure that students will have no prior access to any of the test items. It was included in part because of its greater security and in part because of a high correlation which had on one occasion been secured between performance on it and evaluations of writing samples (2:1&4). Following is a list of the specific test forms employed on the successive testing occasions:

Testing Date	COOP	CEEB
Sept., 1963	1A	KPL1
January, 1964	1B	KPL2
May, 1964	1A	KPL1
May, 1965	1A	HB02

The COOP contains 90 items--30 on Effectiveness and 60 on Mechanics. Total time limit is 40 minutes. The CEEB contains from 100 to 110 items and has a total working time of 60 minutes--20 minutes recommended for each of three sections. From test form to test form the elements tested by the CEEB vary somewhat. Representative elements include paragraph organization, construction shifts, sentence correctness, and usage. The various forms of the test are regarded as equivalent but not parallel.

The theme was a paper written within a two-hour period on a single topic provided by the investigators. Students were urged to remain for the full two-hour period, though they were allowed to leave after an hour and twenty minutes. An explanation of the method for selecting topics, a theme instruction sheet, and the topics used on the various testing dates are included as Appendix A.

Establishing Matched Pairs

For comparing the performance of the two subgroups the investigators used matched pairs instead of the analysis of covariance technique. A discussion of the pros and cons of using the matched pairs approach may be found in Appendix B. The procedure worked out in the following manner.

Enrollment practices at the State College of Iowa made it necessary to select the members of the experimental subgroup--the students who would not enroll in freshman composition--before the beginning of the fall, 1963, registration. To accomplish this, the investigators selected an experimental pool of 325 students. Consultation with the Registrar indicated that most students who enroll in September have been accepted by July 1. He provided the investigators with a list of the names of these students as of July 1, 1963. The investigators separated the members of this group by sex, and within each sex, ranked the students from high to low in terms of performance, as indicated by standard score on the English section of the American College Testing Program (ACT). The goal was to select a group which would include approximately one-third of the entering freshman class, would contain a ratio between men and women representative of the total freshman class, and would reflect the range of performance of that class on the English section of the ACT. The total number of students who had applied by July 1, 1963, was 929 (361 male, 39%--and 568 female, 61%). The experimental pool of 325 (38% male and 62% female) constituted about thirty-five per cent of the total group.

To obtain this group of 325, the investigators assigned a three-digit number to each of the names on the Registrar's list. By use of a table of random numbers, they then selected thirty-five per cent of the students of each sex at each score level. The resulting list was screened by the Registrar to eliminate those whose college programs would terminate before the end of the two-year period of the experiment or who were planning to attend college only part-time. Any student eliminated by this screening procedure was replaced on a random basis by another student of the same sex and ACT score. The 325 students thus identified constituted the experimental pool.

The Dean of Instruction mailed a personal letter to all members of this pool informing them that because they had been selected for a special investigation they should not enroll in freshman composition during their first two years of college. He also invited them to write to the investigators if they had any questions about their participation or if they desired further information. It was expected that this letter would encourage the students to cooperate. Very few students made inquiries and none asked to be taken out of the group. Consequently members of the experimental subgroup were not volunteers; they were in fact selected by the investigators according to the procedure outlined here.

During the orientation period which preceded the beginning of instruction in the fall, 1963, all entering students were given tests which included the Project instruments. After the scores on these tests became available, members of the experimental pool were paired with students enrolled in the freshman composition sequence. Students were paired on the basis of sex, theme score, age, and a score representing combined performance on the CEEB and COOP.

The matching process may be illustrated from actual data for three pairs of students. The numerals represent, in order, the student's sex (1 for male, 2 for female), total theme score (sum of two ratings), age in years, and combined objective test score.

Subgroup	Sex	Total Theme Score	Age	Sum of Two Objective Test Scores
Experimental	1	3	18	77
Control	1	3	18	76
Experimental	1	12	18	140
Control	1	12	17	139
Experimental	2	7	17	101
Control	2	7	17	103

No students were matched unless they were of the same sex, had the same total theme score, were within one year of each other in age, and were within three points of one another on combined objective test scores.

The combining of the scores of the two objective tests was accomplished by using the CEEB Standard Rating and the COOP Converted Score, transforming each into a new standard score on a scale having a mean of 50 and a standard deviation of 10, and adding the two resulting transformed scores. Whenever more than one potential control student qualified as a suitable match for a given member of the experimental pool, selection was by a random procedure. The ratio between the number of students in the experimental pool and the number in the control pool was approximately one to three.

Theme Evaluation

Themes were evaluated by teams selected by Fred Godshalk, Chairman of Test Development in the Humanities at the Educational Testing Service, from the pool of readers used by the Educational

Testing Service in its theme-reading program. These teams were used because of their wide experience with theme reading and because many of the same readers would be used on successive scoring occasions.

The ETS readers were accustomed to a 4-point scale. The SCI investigators preferred a 9-point scale. The goal was to employ a scoring scale which would permit the separation of the themes into a reasonable number of quality levels without presenting the evaluators with so many rating categories that undue time would be consumed in pondering fine distinctions. A compromise was adopted: a 9-point scale (1 to 9) with emphasis on 2, 4, 6, and 8.

When Mr. Godshalk communicated his standards to the readers, they were asked to think of the normal curve as split in the middle, with each segment so created split again halfway between the median and the extreme. This created four categories: much below average, below average; above average, much above average. It did not provide specifically for the average rank. Readers, already accustomed to the four-point scale, found it easy to use 2, 4, 6, and 8 as their main grades, but they were able also to use the odd numbers whenever it seemed that a particular paper had some characteristic requiring a grade between two of the even numbers. Since each paper was read by two readers and the ratings summed, the total possible range of scores for a single paper was from 2 to 18. An explanation of the reading procedure is given in Appendix C.

It is recognized that the validity of these evaluations depends upon the degree to which Mr. Godshalk's judgment of student writing, as modified by discussion with the readers, is sound. Mr. Godshalk has an unusually wide background in evaluating the writing of college-bound high school seniors (3). The readers were from a variety of geographical backgrounds and a wide range of educational institutions. Mr. Godshalk has for years supervised groups of readers like these; the readers have worked together as teams in just such reading situations. Though neither Mr. Cowley nor Mr. Jewell consistently compared their evaluation of sample themes with that of the groups, when they did, there was no pronounced disparity between their ratings and those of the readers. In the judgment of the investigators, the validity of theme evaluations is as high as it is possible to achieve in a project of this sort.

RESULTS

September, 1963, Test Performance of Original and Persisting Subgroups

Table I contains basic information regarding the entire entering freshman class at the State College of Iowa for the academic year 1963-64. None of the information in the table involves student performance after September, 1963. The first line of the table shows the performance for the essentially complete class of new freshmen (N=910) on seven measures. Each successive line of the table represents a specified subgroup of the total group of 910. The data in the table thus permit an examination of the extent to which the persisting experimental and control subgroups, composed of matched pairs of students, remain representative of the parent group.

Line two is of interest as it reveals the extent to which the representativeness of the samples originally identified in the summer of 1963 was retained after the actual enrollment of students in September, 1963. Of the 325 individuals originally drawn for the experimental subgroup in July, 1963, 284 matriculated. Comparison of lines one and two suggests that the basic data for the 284 members of the experimental pool agreed closely with the data for the total freshman class. This close agreement is noted on each of the seven measures. For example, the Cooperative English Tests: English Expression (COOP) converted score mean was 160.49 for the experimental pool and 160.09 for the total class. Thus the goal of the investigators--to select an experimental pool which would be representative of the actual entering freshman class--was achieved. The only aspect in which a noticeable difference exists between the experimental pool and the total class is in the slightly smaller percentage of males found in the experimental pool.

Line three contains the data for the 210 members of the experimental pool who were paired with control students. Line four shows the information for the 210 control students. It will be noted from lines three and four that the experimentals and the controls, as subgroups, were closely matched. On the COOP the means were 160.75 and 160.95 respectively. The variability was similarly close; S. D.'s were 7.99 and 7.63. The means can be compared to the mean of 160.09 for the total freshman class displayed in line one.

Lines 5 and 6 of Table I present the September, 1963, information for complete sets of matched pairs who finished the fall semester 1963-64 with all data available. Again using the COOP

TABLE I

ACHIEVEMENT AS OF SEPTEMBER, 1963, OF THE ENTIRE ENTERING 1963-64 FRESHMAN CLASS AND OF VARIOUS PERSISTING SUBGROUPS OF THAT CLASS

% Men	Entire Class and Subgroups	N	ACT English Stan.Score		ACT Composite Stan.Score		%-ile Rank in H. S. Class		SUI Reading Raw Score		CEEB Eng. Composition Stan.Rating		COOP Eng. Exp.Converted Score		Sept.Theme Total Rating	
			Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.
40.3	All New Freshmen, Sept., 1963	910	22.36	3.41	22.91	3.42	74.73	16.45	48.95	8.94	492.00	86.69	160.09	9.29	8.65	2.24
37.3	Members of Experimental Pool who Actually Enrolled	284	22.51	3.38	22.80	3.39	74.96	16.47 (N=271)	48.73	9.73	494.38	83.19	160.49	8.34	8.49	2.24
36.2	Members of Experimental Pool who were Matched--Experimental Group	210	22.58	3.29	22.74	3.48	75.11	16.46 (N=199)	48.27	10.21	494.35	75.22	160.75	7.99	8.51	2.12
36.2	Control Students who were Matched--Control Group	210	22.57	3.11	23.16	3.43	74.90	16.72 (N=203)	48.38	9.33	491.49	78.48	160.95	7.63	8.48	2.11
32.5	Experimental Group, Jan., 1964	166	22.63	3.39	22.70	3.33	73.43	21.54	48.52	9.57	496.69	74.06	161.28	7.98	8.51	2.06
32.5	Control Group, Jan., 1964	166	22.81	3.15	23.08	3.36	73.36	21.32	47.91	8.62	495.11	78.61	161.43	7.41	8.51	2.07
26.5	Experimental Group, May, 1964	113	22.86	3.37	22.76	3.24	76.56	16.92	48.25	9.97	503.62	67.42	161.48	7.47	8.49	2.04
26.5	Control Group, May, 1964	113	23.01	2.83	23.16	3.09	73.55	21.56	47.78	7.87	498.17	71.53	161.96	6.84	8.49	2.02
19.4	Experimental Group, May, 1965	31	23.45	3.17	23.74	2.57	80.55	14.64	51.74	8.76	515.45	66.67	162.81	6.34	10.23	2.62
19.4	Control Group, May, 1965	31	23.84	2.29	24.26	2.55	80.29	14.22	49.65	7.43	515.90	64.84	162.65	5.84	10.19	2.10

as an example, it will be noted that the experimental subgroup mean was 161.28 and the control subgroup mean was 161.43. The slight selectivity among persisting students as compared to the entering students is seen here; the two subgroups which finished a full semester of college obtained COOP means in September which were slightly higher than the means of 160.09 of the total entering freshman class.

The table shows that at the end of the first full academic year complete data were available for 113 matched pairs of students. The continuing representativeness is reflected by the fact that these 113 pairs had, in September, 1963, means of 161.48 for the experimentals and 161.96 for the controls on the COOP as compared with the mean of the total freshman class (September, 1963) of 160.09.

At the end of two complete academic years, the number of remaining matched pairs was 31. Their means on the COOP in September, 1963, were 162.81 and 162.65. It will be noted that these means are appreciably higher than the mean of 160.09 for the parent group of 910. The factor of selectivity is thus apparent in the somewhat higher means these students achieved in the fall, 1963, testing. It must be remembered that the test performances reported in the table are performances at the beginning of the freshman college year, 1963.

Criterion Scores--September, 1963, through May, 1965

Criterion testing was done at three times--end of first semester, end of second semester, and end of fourth semester. The numbers of matched pairs were, respectively, 166, 113, and 31. Basic comparisons of test performance for each of these three sets of experimental and control students are given: within subgroups between beginning and final means, and between subgroups on ending means only. The appropriate means, standard deviations, r's, and t's will be displayed in the tables.

Table II deals primarily with the criterion tests: the COOP, the CEEB, and a theme. This table presents a compact picture of performance in terms of means and standard deviations on the various criterion facts on the various testing occasions. (The key comparisons and analyses of the data are shown in Tables III-XI.)

Whereas in Table I all test scores were those available in September, 1963, Table II presents the performance of persisting matched pairs at four successive testing periods beginning with September, 1963. Examination of this table will reveal the

TABLE II

PERFORMANCE OF AVAILABLE MATCHED PAIRS OF STUDENTS ON THREE CRITERION MEASURES AT BEGINNING, MIDDLE, AND END OF FIRST YEAR AND END OF SECOND YEAR OF COLLEGE

Subgroup	N	Time of Testing	Cooperative English Tests: English Expression (1960) Converted Score Mean	S.D.	CEEB English Composition Test Standard Rating Mean	S.D.	Theme: Sum of Two Ratings Mean	S.D.
Experimental	210	Sept., 1963	160.75	7.99	494.35	75.22	8.51	2.12
Control	210	"	160.95	7.63	491.49	78.48	8.48	2.11
Experimental	166	Jan., 1964	165.06	7.59	498.12	84.33	9.15	2.47 (N=162)
Control	166	"	165.65	6.14	513.68	80.05	9.20	2.32 (N=164)
Experimental	113	May, 1964	166.08	8.08	537.16	78.29	9.79	2.04
Control	113	"	165.42	7.60	524.46	80.61	9.56	2.02
Experimental	31	May, 1965	168.29	7.45	540.55	72.72	10.23	2.62
Control	31	"	168.06	5.97	551.13	81.61	10.19	2.10

---

differences in performance on the criterion measures for the two subgroups at the beginning of the fall semester, 1963-64; at the end of the fall semester 1963-64; at the end of the spring semester, 1963-64; and at the end of the spring semester, 1964-65. The experimental students did not receive instruction similar to that given in freshman composition (English I, English II). The control students did. Therefore, the data in this table permit the key comparisons of the project: those between the performance of the experimental and control subgroups on the criterion measures at successive points in their college careers.

Comparison of Criterion Scores--Sample Available January, 1964

Table III presents the data on the performance on the COOP of the 166 matched pairs available at the end of the fall semester, January, 1964. Two types of comparisons are presented: the change within each subgroup from September to January and the comparison between subgroups on the January performance.

For the experimental subgroup the test mean in January was 165.06, 3.78 higher than the test mean in September. The resulting t-ratio of 8.66 (df=165) was significant, suggesting that the change in means was greater than could be attributed to chance factors. A similar analysis for the control subgroup shows that the change in mean from 161.43 to 165.65 was also significant (t-ratio=10.45).

When the experimental subgroup is compared with the control subgroup on January test performance, the means differed by only .59. The resulting t-ratio, 1.15, fell short of significance. The hypothesis of equal performance after an equal length of college experience was thus sustained in regard to performance on the COOP at the end of the first semester.

The data in Table IV show the performance on the CEEB of the 166 matched pairs available at the end of the first semester, January, 1964. As with the data for the COOP presented in Table III, the data in Table IV permit an analysis of the change within each subgroup from September to January and a comparison of the January performances of the two subgroups.

For the experimental subgroup, the January mean was 1.43 higher than the September mean (498.12 minus 496.69). This mean change was not significant; the t-ratio was .25. For the control subgroup the mean gain was 18.57, and this was significant (t equals 4.18). Thus the control students advanced more than did the experimental students on CEEB during the fall semester.

TABLE III

THE PERFORMANCE OF 166 MATCHED PAIRS OF STUDENTS ON THE COOPERATIVE ENGLISH TESTS: ENGLISH EXPRESSION AT THE BEGINNING AND AT THE END OF THE FIRST SEMESTER OF COLLEGE; DIFFERENCES IN MEANS, AND t-RATIOS

Subgroup	N	Cooperative English Converted Score Sept. 1963 Mean	Sept. 1963 S.D.	Jan. 1964 Mean	Jan. 1964 S.D.	Difference In Means: Jan. Minus September	r	t-Ratio (df=165)	Difference In Jan.Means: Control Minus Exp.	r	t-Ratio (df=165)
Experimental	166	161.28	7.98	165.06	7.59	3.78	.74	8.66*	.59	.57	1.15
Control	166	161.43	7.41	165.65	6.14	4.22	.72	10.45*			

*Significant at .05 level (two-tailed test); t of 1.98 or higher required for significance at .05 level.

TABLE IV

PERFORMANCE OF 166 MATCHED PAIRS OF STUDENTS ON THE COLLEGE ENTRANCE EXAMINATION BOARD ENGLISH COMPOSITION TEST AT THE BEGINNING AND AT THE END OF THE FIRST SEMESTER OF COLLEGE; DIFFERENCES IN MEANS, AND t-RATIOS

		College Entrance Examination Board Standard Rating									
		Sept. 1963		Jan. 1964							
Subgroup	N	Mean	S.D.	Mean	S.D.	Difference In Means: Jan. Minus September	r	t-Ratio (df=165)	Difference In Jan.Means: Control Minus Exp.	r	t-Ratio (df=165)
Experimental	166	496.69	74.05	498.12	84.33	1.43	.64	.25	15.56	.59	2.69*
Control	166	495.11	78.61	513.68	80.05	18.57	.74	4.18*			

*Significant at .05 level (two-tailed test); t of 1.98 or higher required for significance at .05 level.

An examination of the January test scores reveals that the mean for the control subgroup is 15.56 points higher than the mean for the experimental subgroup. This difference is significant (t=2.69). This is consistent with the data presented for the within-subgroup gains.

The data for theme performance are presented somewhat differently from the manner in which the objective test data are reported. Braddock *et al.* suggest, and the investigators agree, that theme evaluations can be considered comparable only when three conditions are met: the evaluations are all made on the same occasion, the evaluators are ignorant of the time of writing, and the evaluators do not know which papers were written by the experimental students and which were written by the control students (1:10-11). Inasmuch as the experimental procedures required September themes to be evaluated in October, 1963, and the January, 1964, themes to be evaluated in May, 1964, the September and January ratings are not comparable. Consequently, only January, 1964, performance was analyzed. As Table V shows, the mean for the 162 experimental students is 9.15 and the mean for the 164 control students is 9.20. The difference of .05 is so small that it could easily be attributed to chance; *i.e.* the difference is not significant (t=.22).

Comparison of Criterion Scores--Sample Available May, 1964

Whereas Tables III, IV, and V were concerned with the first semester of college, Tables VI, VII, and VIII present evidence for the first and the second semesters. At the end of the second semester, full data were available for 113 of the 166 matched pairs for whom full data were available at the end of the fall semester.

Table VI contains data obtained from the administration of the COOP on three occasions: September, 1963; January, 1964; and May, 1964. The first of the three parts of the table shows the facts for September, 1963, and January, 1964. As with Tables III and IV, the difference between the September and January performances within each subgroup and the difference in January performance between the two subgroups are shown.

The change in mean during the first semester was 3.64 for the experimental subgroup and 3.89 for the control subgroup. For each subgroup, the mean gain was significant: t-ratios were 6.75 and 7.92. A comparison of the January means for the two subgroups revealed an advantage of .73 for the controls. This mean difference at the end of the semester was not significant (t=1.25).

TABLE V

PERFORMANCE OF APPROXIMATELY 166 MATCHED PAIRS OF STUDENTS ON THE SUM OF TWO THEME RATINGS AT THE END OF THE FIRST SEMESTER OF COLLEGE; DIFFERENCE IN MEANS, AND t-RATIO

Subgroup	N*	Mean Total Theme Rating	Standard Deviation	Difference in Jan. Means: Control Minus Experimental	r	t-Ratio (df=about 160)
Experimental	162	9.15	2.45	.05	.23	.22
Control	164	9.20	2.33			

*The N's are slightly less than 166 owing to apparent misplacement of data during processing. t of 1.98 or higher required for significance at .05 level.

TABLE VI

THE PERFORMANCE OF 113 MATCHED PAIRS OF STUDENTS ON THE COOPERATIVE ENGLISH TESTS: ENGLISH EXPRESSION AT THE BEGINNING OF THE FIRST SEMESTER, AT THE END OF THE FIRST SEMESTER, AND AT THE END OF THE SECOND SEMESTER; DIFFERENCES IN MEANS, AND t-RATIOS

Subgroup	N	Cooperative English Converted Scores				Difference In Means: Jan. Minus September	r	t-Ratio (df=112)	Difference In Jan.Means: Control Minus Exp.	r	t-Ratio (df=112)
		Sept. 1963		Jan. 1964							
		Mean	S.D.	Mean	S.D.						
Experimental	113	161.48	7.47	165.12	7.06	3.64	.69	6.75*	+.73	.53	1.25
Control	113	161.96	6.84	165.85	5.45	3.89	.66	7.92*			

Subgroup	N	Cooperative English Converted Scores				Difference In Means: May Minus January	r	t-Ratio (df=112)	Difference In May Means: Control Minus Exp.	r	t-Ratio (df=112)
		Jan. 1964		May 1964							
		Mean	S.D.	Mean	S.D.						
Experimental	113	165.12	7.06	166.08	8.08	.96	.75	1.88	-.66	.49	.88
Control	113	165.85	5.45	165.42	7.60	-.43	.70	.84			

Subgroup	N	Cooperative English Converted Scores				Difference In Means: May Minus September	r	t-Ratio (df=112)	Difference In May Means: Control Minus Exp.	r	t-Ratio (df=112)
		Sept. 1963		May 1964							
		Mean	S.D.	Mean	S.D.						
Experimental	113	161.48	7.47	166.08	8.08	4.60	.72	8.36*	-.66	.49	.88
Control	113	161.96	6.84	165.42	7.60	3.46	.61	5.74*			

*Significant at .05 level (two-tailed test); t of 1.98 or higher required for significance at .05 level.

It is useful at this point to consider certain aspects of significant difference in this setting. In the preceding paragraph it was noted that during the first semester the COOP mean gain scores were significant, whereas the January COOP scores of the experimental subgroup and the Control subgroup did not differ significantly. Specifically, a difference of .73 was not significant. What kind of a difference between the two subgroups on the January COOP testing would have been large enough to be significant? This can be readily estimated. The obtained t value was 1.25; the value needed for significance was 1.98, i.e., the obtained t was 63% as large as the needed t. It follows that the obtained mean difference was 63% as large as the mean difference needed for significance. Therefore, if instead of the actual difference of .73 the obtained mean difference had been 1.16, the difference would have been significant at the .05 level. What does this mean in terms of performance on the COOP test? If on this 90-item test each student in one of the subgroups had given one or two more correct answers than did his counterpart in the other subgroup, and if standard deviations and correlations remained about the same as those reported in Table VI, a significant difference would have occurred.

The second section of the table gives data for the second semester--the period between January, 1964, and May, 1964. The experimental subgroup had a mean gain in this period of .96, and this was not significant (t=1.88). The control subgroup showed a slight decrease in mean test score, .43; this decrease was not significant (t=.84). On the May, 1964, testing the experimental subgroup mean was .66 higher than the control subgroup mean. This was not a significant difference (t=.88; df=112).

During the first semester, then, both subgroups made a significant improvement in performance on the COOP; during the second semester neither subgroup did.

The third section of Table VI presents the performance of the two subgroups at the beginning of the first semester and at the end of the second semester, 1963-64. Over this nine-month period, the experimental subgroup showed an increase in mean test score from 161.48 to 166.08. This mean gain of 4.60 was significant (t=8.36). The control subgroup advanced in test mean from 161.96 to 165.42. This mean gain of 3.46 was also significant (t=5.74). For each subgroup, the significant improvement during the year was actually achieved during the first semester. Apparently the experience reflected in the observed change in test scores occurred during the first semester, and no experience during the second semester resulted in a significant additional change.

In the analyses stemming from Table VI, correlations between sets of test scores are utilized. It may be noted that repeated testing of the same individuals showed r's of the order of .70. The correlations across matched pairs in May were approximately .50. At the beginning of the experiment the correlation across matched pairs was .76. The time interval between the testings was nine months. The correlation data tend to confirm the idea that the original matching was reasonably satisfactory.

Table VII, based upon the CEEB, is similar to Table VI, which dealt with the COOP. The facts for the first semester are in the upper section of Table VII. For the experimental subgroup, the September, 1963, to January, 1964, mean change was minus 2.43--a slight decline from 503.62 to 501.19. The control subgroup showed a change in mean from 498.17 (September) to 517.77 (January). The increase of 19.60 was significant (t=3.54). The January means for the two subgroups differ by 16.58, and the t-value of 2.36 indicates a significant advantage in favor of the control subgroup.

The January and May CEEB test data appear in the middle section of Table VII. It is noteworthy that, for the change scores, the findings are a reversal of those just examined for the first semester. During the second semester, the experimental subgroup improved significantly, whereas the control subgroup did not. The respective changes in mean test scores were 35.97 for the experimental subgroup and 6.69 for the control subgroup. In the comparison of the May test score means, the experimental subgroup was 12.70 points higher than the control subgroup. This differential was not quite significant (t=1.65; 1.98 required to achieve significance). For the second semester, then, the experimental subgroup started with a lower mean test rating than the control subgroup and finished with a higher mean test rating.

By using the upper and middle sections of Table VII, it is possible to explore some of the requirements for a significant difference on the CEEB. Let us use as an example the difference in means--control minus experimental--for January, 1964, and for May, 1964. The difference between subgroups in January, 16.58, was significant; the difference in May, 12.70, was not significant. Thus, in terms of the standard deviations and correlations involved, a between-subgroup mean difference of about 14 or 15 was required for significance at the .05 level. For the CEEB test an increase of one raw score point--one more question right--is typically associated with an increase of about six standard rating points. Thus if each member of one of the subgroups had had two or three more correct responses than did his counterpart in the other subgroup, the resulting subgroup means

TABLE VII

THE PERFORMANCE OF 113 MATCHED PAIRS OF STUDENTS ON THE COLLEGE ENTRANCE EXAMINATION BOARD ENGLISH COMPOSITION TEST AT THE BEGINNING OF THE FIRST SEMESTER, AT THE END OF THE FIRST SEMESTER, AND AT THE END OF THE SECOND SEMESTER; DIFFERENCES IN MEANS, AND t-RATIOS

		College Entrance Examination Board Standard Rating									
		Sept. 1963		Jan. 1964							
Subgroup	N	Mean	S.D.	Mean	S.D.	Difference In Means: Jan. Minus September	r	t-Ratio (df=112)	Difference In Jan.Means: Control Minus Exp.	r	t-Ratio (df=112)
Experimental	113	503.62	67.42	501.19	80.35	-2.43	.58	.38	+16.58	.54	2.36*
Control	113	498.17	71.53	517.77	75.32	19.60	.68	3.54*			

		College Entrance Examination Board Standard Rating									
		Jan. 1964		May 1964							
Subgroup	N	Mean	S.D.	Mean	S.D.	Difference In Means: May Minus January	r	t-Ratio (df=112)	Difference In May Means: Control Minus Exp.	r	t-Ratio (df=112)
Experimental	113	501.19	80.35	537.16	78.29	35.97	.70	6.22*	-12.70	.47	1.65
Control	113	517.77	75.32	524.46	80.61	6.69	.54	.95			

		College Entrance Examination Board Standard Rating									
		Sept. 1963		May 1964							
Subgroup	N	Mean	S.D.	Mean	S.D.	Difference In Means: May Minus September	r	t-Ratio (df=112)	Difference In May Means: Control Minus Exp.	r	t-Ratio (df=112)
Experimental	113	503.62	67.42	537.16	78.29	33.54	.65	5.77*	-12.70	.47	1.65
Control	113	498.17	71.53	524.46	86.61	26.29	.63	4.14*			

*Significant at .05 level (two-tailed test); t-ratio of 1.98 or higher required for significance.

would have differed significantly. The number of items on the several forms of the CEEB ranges from 100 to 110.

The lower section of Table VII compares the September performance with the May performance. Both subgroups experienced a significant gain in mean test performance. The mean change for the experimental subgroup was 33.54, and for the control subgroup, 26.29. The associated t-ratios were 5.77 and 4.14. This significant increment during the academic year took place during the first semester for the control subgroup and during the second semester for the experimental subgroup. These results mark an interesting contrast to the results on the COOP, in which both subgroups made a significant gain during the first semester, and no significant additional gain during the second semester. It should be remembered that the experimental students did not receive instruction in freshman composition during either semester, whereas the control students had such instruction both semesters.

Table VIII contains the results of the administration of a theme in May, 1964. The theme rating for each student was the sum of two independent evaluations. It will be seen that the means for the experimental subgroup and the control subgroup are quite similar: 9.79 and 9.56. The difference of .23 may be interpreted in terms of a t-ratio of .86, which is not significant (t of 1.98 required at .05 level).

It is appropriate to attempt some interpretation of the fact that the obtained t value of .86 was not significant, but one of 1.98 would have been. Since a t value 2.29 times as large as the obtained one was required for significance, the mean difference would have had to be 2.29 times as large. That is, for significance, the between subgroups mean difference would have needed to be .53 instead of the .23 obtained. Thus, if, on the theme in May, 1964, (about) 55 members of one subgroup had the same theme score as their counterparts in the other subgroup, but the other 55 members of the one subgroup scored one point higher than their paired counterparts, the resulting subgroup means would have differed significantly. This illustrative analysis is in terms of a level of confidence of .05 and of standard deviations and correlations similar to those in Table VIII. The possible range of Total Theme score was from 2 to 18.

It is interesting to note the theme ratings of these same 110 matched pairs at the end of the first semester. The papers written in January and in May were read at the same time by the same team of readers; the readers did not know which papers were written by the experimental and which by the control students,

TABLE VIII

THE PERFORMANCE OF 110 MATCHED PAIRS OF STUDENTS ON THE TOTAL OF TWO THEME RATINGS AT THE END OF THE SECOND SEMESTER; DIFFERENCES IN MEANS, AND t-RATIOS

Subgroup	N*	Theme In May, 1964		Difference In Means on Theme: Control Minus Exp.	r	t-Ratio (df=109)
		Mean	S.D.			
Experimental	110	9.79	2.04	-.23	.03	.86
Control	110	9.56	2.02			

*The N's are slightly less than 113 because of apparent misplacement of data during processing.

and they did not know which topic was used in January and which in May. In the light of these conditions, we can in this instance compare theme performance at different testing occasions. The January, 1964, results were as follows:

N		Mean	S.D.	r
110	Experimental	9.25	2.40	.23
112	Control	9.37	2.31	

The January means of 9.25 and 9.37 did not differ significantly (t=.43; df=109). The gain during the second semester for the experimental subgroup was .54. The January-May correlation was .37, and the t-ratio of 1.80 was not significant. The gain for the control subgroup during the second semester was .19. The January-May correlation was also .37, and the t-ratio of .67 was not significant (df=111).

Comparison of Criterion Scores--Sample Available May, 1965

Tables IX, X, and XI show the data through the first two years of college--the total interval covered in the present study.* They present evidence from four testing occasions for the 31 matched pairs of students for whom full data were available in May, 1965.

Table IX is based upon student performance on the COOP. The first set of facts in Table IX is for the first semester of the freshman year (September, 1963-January, 1964). The experimental subgroup had a gain of 1.96 over the semester, and this was significant (t=2.14). The control subgroup showed a mean gain of 5.12. This, also, was significant (t=7.06). The January means were 164.77 and 167.77. This difference of 3.00 in favor of the control subgroup was significant (t=2.65).

During the second semester of the freshman year, the experimental subgroup showed a significant gain, while the control subgroup did not. The data appear in the second section of Table IX. The change in mean test performance was 3.04 (t=3.15) for the experimental students and .58 (t=.81) for the control students.

---

*Research Project 3177, an extension of the present investigation, provides for testing the entire senior class in the spring of 1967. Some of the students who have been involved in the present study will be among those tested at that time.

TABLE IX

THE PERFORMANCE OF 31 MATCHED PAIRS OF STUDENTS ON THE COOPERATIVE ENGLISH TESTS: ENGLISH EXPRESSION AT THE BEGINNING OF THE FIRST SEMESTER, AT THE END OF THE FIRST SEMESTER, AT THE END OF THE SECOND SEMESTER, AND AT THE END OF THE FOURTH SEMESTER; DIFFERENCES IN MEANS, AND t-RATIOS

Sub-group	N	Cooperative English Converted Scores				Diff.In Means: Jan. Minus Sept.	r	t-Ratio (df=30)	Difference In Jan. Means: Control Minus Exp.	r	t-Ratio (df=30)
		Sept. 1963		Jan. 1964							
		Mean	S.D.	Mean	S.D.						
Experimental	31	162.81	6.34	164.77	7.13	1.96	.72	2.14*	+3.00	.53	2.65*
Control	31	162.65	5.84	167.77	5.56	5.12	.75	7.06*			

Sub-group	N	Cooperative English Converted Scores				Diff.In Means: May Minus Jan.	r	t-Ratio (df=30)	Difference In May Means: Control Minus Exp.	r	t-Ratio (df=30)
		Jan. 1964		May 1964							
		Mean	S.D.	Mean	S.D.						
Experimental	31	164.77	7.13	167.81	7.25	3.04	.72	3.15*	+.54	.42	.42
Control	31	167.77	5.56	168.35	5.70	.58	.75	.81			

Sub-group	N	Cooperative English Converted Scores				Diff.In Means: May Minus Sept.	r	t-Ratio (df=30)	Difference In May Means: Control Minus Exp.	r	t-Ratio (df=30)
		Sept. 1963		May 1964							
		Mean	S.D.	Mean	S.D.						
Experimental	31	162.81	6.34	167.81	7.25	5.00	.61	4.60*	+.54	.42	.42
Control	31	162.65	5.84	168.35	5.70	5.70	.74	7.62*			

Sub-group	N	Cooperative English Converted Scores				Diff.In Means: May Minus May	r	t-Ratio (df=30)	Difference In May '65 Means: Control Minus Exp.	r	t-Ratio (df=30)
		May 1964		May 1965							
		Mean	S.D.	Mean	S.D.						
Experimental	31	167.81	7.25	168.29	7.45	.48	.72	.49	.23	.47	.18
Control	31	168.35	5.70	168.06	5.97	-.29	.76	.40			

Sub-group	N	Cooperative English Converted Scores				Diff.In Means: May Minus Sept.	r	t-Ratio (df=30)	Difference In May Means: Control Minus Exp.	r	t-Ratio (df=30)
		Sept. 1963		May 1965							
		Mean	S.D.	Mean	S.D.						
Experimental	31	162.81	6.34	168.29	7.45	5.48	.75	6.12*	.23	.47	.18
Control	31	162.65	5.84	168.06	5.97	5.41	.72	6.81*			

*t-ratio of 2.04 or higher required for significance at .05 level (two-tailed test)

The effect of the gain of the control subgroup during the first semester and of the experimental subgroup during the second semester was that by May, 1964, the means of 167.81 and 168.35 did not differ significantly (t=.42).

The third section of Table IX covers the entire freshman year (September, 1963-May, 1964). During the nine months, the experimental subgroup mean increased from 162.81 to 167.81. The corresponding increase for the control subgroup was from 162.65 to 168.35. The gains, 5.00 and 5.70 respectively, are both significant. The similarity between the two subgroups at the beginning and at the end of the academic year is noteworthy.

The test results of May, 1964, and May, 1965, are presented in section four of Table IX. The end-of-year performances as freshmen and sophomores were strikingly similar: for the experimental subgroup, 167.81 and 168.29; for the control subgroup, 168.35 and 168.06. The small changes were, of course, not significant. The between-subgroup difference in May, 1965, was .23 (not significant--t=.18).

The final section of Table IX contains test data obtained in September, 1963, and in May, 1965. Over the two academic years both the experimental subgroup and the control subgroup gained significantly. The mean gains were 5.48 for the experimentals, and 5.41 for the controls.

The notable fact which may be derived from Table IX is that the students' performance on COOP did not improve during the second year of the interval (May, 1964-May, 1965). For the control subgroup, the significant improvement in performance occurred during the first semester of the first year; for the experimental subgroup the significant improvement came during the second semester of the first year. A plausible explanation of these data would be that the performance of the control group in January, 1964, is the result of instruction while the performance of the experimental group in May, 1964, is the result of maturation. Further, one more year of maturation had no visible effect on either group's performance on COOP.

Table X covers the same span of time and performances as does Table IX--the beginning, middle, and end of the freshman year and the end of the sophomore year. The test scores were obtained from administrations of the CEEB. For the first semester of the freshman year, the experimental subgroup had an initial mean of 515.45 and an ending mean of 519.13. This gain of 3.68 was not significant (t=.33; df=30). The control subgroup advanced from a September mean of 515.90 to a January mean of

TABLE X

THE PERFORMANCE OF 31 MATCHED PAIRS OF STUDENTS ON THE COLLEGE ENTRANCE EXAMINATION BOARD ENGLISH COMPOSITION TEST AT THE BEGINNING OF THE FIRST SEMESTER, AT THE END OF THE FIRST SEMESTER, AT THE END OF THE SECOND SEMESTER, AND AT THE END OF THE FOURTH SEMESTER; DIFFERENCE IN MEANS, AND t-RATIOS

Sub-group	N	College Entrance Examination Board Standard Rating				Diff.In Means: Jan. Minus Sept.	r	t-Ratio (df= 30)	Difference In Jan. Means: Control Minus Exp.	r	t-Ratio (df= 30)
		Sept. 1963		Jan. 1964							
		Mean	S.D.	Mean	S.D.						
Experi-mental	31	515.45	66.67	519.13	72.21	3.68	.61	.33	+18.77	.67	1.64
Control	31	515.90	64.84	537.90	82.64	22.00	.74	2.20*			

Sub-group	N	College Entrance Examination Board Standard Rating				Diff.In Means: May Minus Jan.	r	t-Ratio (df= 30)	Difference In May Means: Control Minus Exp.	r	t-Ratio (df= 30)
		Jan. 1964		May 1964							
		Mean	S.D.	Mean	S.D.						
Experi-mental	31	519.13	72.21	558.87	74.60	39.74	.74	4.18*	-1.03	.62	.08
Control	31	537.90	82.64	557.84	82.43	19.94	.61	1.52			

Sub-group	N	College Entrance Examination Board Standard Rating				Diff.In Means: May Minus Sept.	r	t-Ratio (df= 30)	Difference In May Means: Control Minus Exp.	r	t-Ratio (df= 30)
		Sept. 1963		May 1964							
		Mean	S.D.	Mean	S.D.						
Experi-mental	31	515.45	66.67	558.87	74.60	43.42	.75	4.79*	-1.03	.62	.08
Control	31	515.90	64.84	557.84	82.43	41.94	.62	3.53*			

Sub-group	N	College Entrance Examination Board Standard Rating				Diff.In Means: May Minus May	r	t-Ratio (df= 30)	Difference In May '65 Means: Control Minus Exp.	r	t-Ratio (df= 30)
		May 1964		May 1965							
		Mean	S.D.	Mean	S.D.						
Experi-mental	31	558.87	74.60	540.55	72.72	-18.32	.83	2.37*	+10.58	.44	.72
Control	31	557.84	82.43	551.13	81.61	-6.71	.65	.54			

Sub-group	N	College Entrance Examination Board Standard Rating				Diff.In Means: May Minus Sept.	r	t-Ratio (df= 30)	Difference In May Means: Control Minus Exp.	r	t-Ratio (df= 30)
		Sept. 1963		May 1965							
		Mean	S.D.	Mean	S.D.						
Experi-mental	31	515.45	66.67	540.55	72.72	25.10	.54	2.08*	+10.58	.44	.72
Control	31	515.90	64.84	551.13	81.61	35.23	.61	2.95*			

*t-ratio of 2.04 or higher required for significance at .05 level (two-tailed test)

537.90. This 22-point gain was significant (t=2.20; df=30). The January, 1964, mean for the control subgroup was 18.77 higher than the mean for the experimental subgroup, and this difference was not significant (t=1.64; df=30).

The evidence for the second semester of the freshman year appears in the second section of Table X. The experimental subgroup had a mean change of 39.74 (558.87 minus 519.13); this was significant as revealed by a t-ratio of 4.18. The control subgroup had a mean change of 19.94 (557.84 minus 537.90); this was not significant in terms of a t-ratio of 1.52. The May, 1964, test means of the two subgroups differed by only 1.03. The t-ratio was .08.

Evidence over the total freshman year is presented in the third section of Table X. Significant gains between September and May are noted for both the experimental subgroup and the control subgroup: 43.42 (experimental) and 41.94 (control). The corresponding t-values were 4.79 and 3.53.

The last two sections of Table X include the test means for May, 1965. The fourth section shows that between May, 1964, and May, 1965, both subgroups declined somewhat in their performance. The negative change of 18.32 for the experimental subgroup was significant (t=2.37; df=30). The negative change of 6.71 for the control subgroup was not significant (t=.54; df=30). The May, 1965, subgroup means were 540.55 for the experimental subgroup and 551.13 for the control subgroup. This difference of 10.58 was not significant (t=.72; df=30).

The final section of Table X covers a span of two academic years. When the September, 1963, CEEB scores are compared with the May, 1965, scores a significant gain is noted for both subgroups. These mean gains were 25.10 (experimental) and 35.23 (control). The associated t-ratios were 2.08 and 2.95.

Table XI shows the facts regarding the theme written at the end of the sophomore year. It will be remembered that the theme of each student was read by two raters and the student's score was the sum of the two independent ratings. The means for the experimental subgroup and the control subgroup were strikingly similar--10.23 and 10.19. One may acquire some notion of the distribution of theme ratings within subgroups from the obtained sigmas of 2.62 and 2.10. The across-subgroups correlation for the 31 matched pairs was .23.

Tables II through XI have compared the performance of the experimental subgroup and the control subgroup on the three

TABLE XI

THE PERFORMANCE OF 31 MATCHED PAIRS OF STUDENTS ON THE TOTAL OF TWO THEME RATINGS AT THE END OF THE FOURTH SEMESTER (MAY, 1965); DIFFERENCES IN MEANS, AND t-RATIOS

Subgroup	N	Theme Ratings Mean	Theme Ratings S.D.	Difference In Means: Control Minus Exp.	r	t-Ratio (df=30)
Experimental	31	10.23	2.62	-.04	.23	.08
Control	31	10.19	2.10			

criterion measures at three testing junctures. At the end of the first semester (January, 1964), there was evidence for 113 of the 166 matched pairs of students; at the end of the fourth semester (May, 1965), there was evidence for 31 of the matched pairs. These are the three main subsamples.

It will be seen from Table XII that on the COOP none of the three main subsamples showed a significant mean difference between subgroups. That is, when comparisons were made at each of the three testing points using the maximum number of available matched pairs, the hypothesis of equality of performance of experimental subgroup and control subgroup was supported.

The evidence was different for CEEB. Here, the end of the first semester marked a superiority for the control subgroup over the experimental subgroup. This significant superiority did not persist through the second semester or the second academic year.

Theme scores were such that in each of the three key comparisons there was no significant difference between subgroup means.

Further inspection of Table XII reveals that for COOP one of the three secondary comparisons yielded a significant mean difference: for 31 matched pairs as of January, 1964, the control subgroup was superior. For CEEB there was also a significant difference in favor of the control subgroup in one of the three secondary comparisons: for 113 matched pairs as of January, 1964. Theme scores did not produce any significant between-subgroup differences in the three secondary comparisons.

INTERCORRELATIONS AMONG VARIABLES

September, 1963--Total Group

Table XIII contains coefficients of correlation between all possible pairs of nine variables for 910 new freshmen at the State College of Iowa at the beginning of the fall semester, 1963. Two of the variables--ACT Composite and Percentile Rank in High School Graduating Class--customarily serve as indices of over-all high school accomplishment and of general potential for college study. The other seven variables are test scores in the area of language arts. These intercorrelation data, together with the related means and standard deviations reported in Table I (p. 21), provide a description of the 1963 SCI freshman class pertinent to an investigation of their writing performance.

The highest correlations were between each of the independent theme ratings and the sum of these two (Total Theme score);

TABLE XII

A SUMMARY COMPARISON OF TEST SCORE MEANS OF EXPERIMENTAL SUBGROUP AND CONTROL SUBGROUP AT THREE TESTING POINTS: INDICATION OF STATISTICALLY SIGNIFICANT DIFFERENCES FOR SPECIFIED SUBSAMPLES

Did one of the two subgroups have a significantly higher mean than the other?

	January 1964			May 1964		May 1965
	N=166	N=113	N=31	N=113	N=31	N=31
Cooperative English Tests: English Expression	No	No	Yes Control	No	No	No
College Entrance Examination Board English Composition Test	Yes Control	Yes Control	No	No	No	No
Theme	No	No	No	No	No	No

TABLE XIII

INTERCORRELATIONS AMONG 9 VARIABLES FOR 910 NEW FRESHMEN ENTERING THE STATE COLLEGE OF IOWA, FALL SEMESTER, 1963-64

Variable	ACT English Stan.Score	ACT Comp. Stan.Score	SUI Reading Raw Sc.	CEEB Eng.Comp. Stan.Rat.	COOP Eng. Converted Score	Theme Rating 1	Theme Rating 2	Sum of Theme Ratings	%-ile Rank H.S.
ACT English Standard Score									
ACT Composite Standard Score	.62								
SUI Reading Raw Score	.45	.68							
CEEB English Comp. Standard Rating	.65	.62	.60						
COOP English Test Converted Score	.59	.45	.38	.59					
Theme Rating Sept. by Reader 1	.35	.21	.21	.33	.26				
Theme Rating Sept. by Reader 2	.35	.25	.25	.37	.29	.35			
Sum of Theme Ratings of Readers 1 & 2	.43	.28	.28	.43	.33	.85	.78		
%-ile Rank in H.S. Graduating Class	.51	.43	.33	.43	.37	.25	.27	.32	

INTERCORRELATIONS AMONG 9 VARIABLES FOR 910 NEW FRESHMEN ENTERING THE STATE COLLEGE OF IOWA FALL SEMESTER, 1963-64

Variables Involved

1 American College Testing Program (ACT) English standard score

2 American College Testing Program (ACT) Composite standard score

3 State University of Iowa (SUI) Reading Comprehension Form A raw score

4 College Entrance Examination Board English Composition Test Form KPL1 standard rating

5 Cooperative English Tests: English Expression Form 1A converted score

6 Theme rating assigned by Reader 1 (September, 1963)

7 Theme rating assigned by Reader 2 (September, 1963)

8 Sum of theme ratings assigned by Readers 1 and 2 (September, 1963)

9 Percentile rank in high school graduating class

the r's were .85 and .78. It is of special interest that the between-readers r was .35.

Which of the three objective tests correlated highest with the Total Theme Rating? Both the ACT English and the CEEB showed an r of .43, whereas for COOP the r was .33. How did the three objective tests correlate with one another? The three intercorrelations were .59, .59, and .65 (the last, ACT English--CEEB). Since the ACT English score was utilized in selecting the members of the experimental pool (basic planning which had to be conducted in the summer before the students entered college), it is of interest to see that ACT correlated reasonably well with the criterion measures. The correlation is all the more noteworthy when one remembers that the ACT battery of tests was in all cases taken at least four months prior to the September administration of the criterion tests, and may have been taken as many as 10 months before.

January, 1964--End of First Semester

Table XIV is an intercorrelation matrix involving 15 variables. The 332 students are the combined experimental and control subgroups completing the first semester of the 1963-64 academic year with complete data. These 332 students were a part of the 910 for whom beginning-of-the-year intercorrelation data for nine variables were presented in Table XIII.

The six additional variables available in January were the January values for CEEB, COOP, Theme Rating one, Theme Rating two, Theme Total, and grade-point average for the fall semester. Of the September variables, ACT English showed the highest relationship with January theme total (r=.48). Another one of the September-January comparisons is between theme total on the two occasions; the obtained r was .40. Still another September-January relationship of interest is for the scores on CEEB; for this, the r was .69. The corresponding figure for COOP was .59. The September measure which showed the highest relationship with first semester grade point average was ACT Composite; the r was .61.

Intercorrelation data for the subgroup of 166 experimental students appears in Table XV. For this subgroup, the January theme total was best predicted by the September ACT English score (r=.55) and CEEB (r=.53). For COOP the corresponding figure was .45. In general, the r's for the 166 experimental students were not markedly different from the r's for the experimental subgroup plus the control subgroup (N=332).

TABLE XIV

INTERCORRELATIONS AMONG 15 VARIABLES FOR 332 STUDENTS (166 MATCHED PAIRS) AT THE END OF THE FALL SEMESTER, 1963-64

Variable	ACT Eng	ACT Comp	SUI Read	CEEB Sept	COOP Eng Sept	Theme Rating September			%-ile Rank H.S.	CEEB Jan.	COOP Eng. Jan.	Theme Rating January			GPA Fall Sem
						1	2	Total				1	2	Total	
ACT English															
ACT Composite	.63														
SUI Reading	.46	.67													
CEEB September	.63	.63	.59												
COOP Eng. Sept.	.50	.39	.29	.45											
Sept. Theme: Reader 1	.40	.28	.25	.38	.29										
Reader 2	.35	.27	.26	.40	.20	.34									
Sum 1 & 2	.46	.33	.31	.48	.31	.85	.78								
%-ile Rank H.S.	.34	.29	.23	.31	.21	.21	.22	.26							
CEEB January	.69	.61	.53	.69	.49	.40	.37	.47	.30						
COOP Eng. Jan.	.63	.53	.43	.60	.59	.36	.36	.44	.32	.67					
Jan. Theme: Reader 1	.42	.33	.32	.34	.25	.28	.25	.33	.20	.40	.36				
Reader 2	.39	.32	.31	.38	.29	.27	.29	.34	.21	.37	.38	.41			
Sum 1 & 2	.48	.39	.37	.43	.32	.33	.32	.40	.25	.46	.44	.85	.82		
GPA Fall Sem	.44	.61	.53	.46	.33	.27	.26	.32	.38	.40	.42	.23	.32	.33	

TABLE XV

INTERCORRELATIONS AMONG 15 VARIABLES FOR 166 EXPERIMENTAL STUDENTS AT THE END OF THE FALL SEMESTER, 1963-64

Variable	ACT Eng	ACT Comp	SUI Read	CEEB Sept	COOP Eng Sept	Theme Rating September 1	Theme Rating September 2	Theme Rating September Total	%-ile Rank H.S.	CEEB Jan.	COOP Eng. Jan.	Theme Rating January 1	Theme Rating January 2	Theme Rating January Total	GPA Fall Sem
ACT English															
ACT Composite	.59														
SUI Reading	.41	.70													
CEEB September	.64	.62	.56												
COOP Eng. Sept.	.67	.52	.39	.63											
Sept. Theme: Reader 1	.42	.26	.20	.35	.38										
Reader 2	.36	.28	.25	.40	.30	.31									
Sum 1 & 2	.48	.33	.28	.46	.43	.83	.78								
%-ile Rank H.S.	.34	.32	.24	.28	.37	.31	.24	.34							
CEEB January	.68	.57	.52	.64	.64	.45	.33	.49	.31						
COOP Eng. Jan.	.62	.52	.38	.58	.74	.33	.32	.40	.31	.66					
Jan. Theme: Reader 1	.45	.36	.34	.39	.37	.31	.22	.33	.25	.43	.37				
Reader 2	.48	.36	.37	.50	.38	.30	.30	.37	.34	.42	.42	.42			
Sum 1 & 2	.55	.43	.42	.53	.45	.36	.32	.32	.35	.50	.47	.86	.82		
GPA Fall Sem	.38	.61	.51	.47	.35	.26	.29	.34	.39	.38	.38	.25	.36	.36	

At the end of the fall semester, grades in freshman composition were available for the 166 control students. From the evidence in Table XVI, it is possible to rate the potency of the various instruments in predicting students' success in the composition course. The following listing shows the correlation between first semester grade in composition and evidence available in September on each of several indexes:

ACT English	.54
ACT Composite	.46
CEEB	.44
COOP	.40
SUI Reading	.37
Percentile Rank in H.S.	.33
Sept. Theme Total	.33

Thus both theme performance in January and the final grade in the first semester freshman composition course are best predicted by the ACT English score. This finding is of interest in that the two January indexes are considered to be direct measures of writing, while ACT English is relatively indirect. This finding for the ACT English score as a predictor is the more noteworthy when one remembers that the ACT test battery was typically taken during the students' senior year in high school.

May, 1964--End of First Year

For the combined subgroups available at the end of the first year of college (N=113 + 113), the intercorrelation matrix involved 21 variables. Table XVII contains the correlation data for selected pairs of variables. In the first section of the table, the r's are for total theme rating in May and each of four September scores. In addition to the coefficients for the combined subgroups, the table also shows the coefficients by subgroup.

For samples in which N=113, a correlation coefficient must be as large as .18 to be significant at the .05 level of confidence. Ten of the twelve correlations between September indexes and May theme total were significant--that is, indicated that in the population sampled the correlation differed from zero by some amount. Among the twelve comparisons, the r's ranged from .10 to .29.

It is interesting to note the relationship between the September scores and the May scores for each of the two objective tests which were repeated. For the combined subgroups the r's were .67 for COOP, and .64 for CEEB.

TABLE XVI

INTERCORRELATIONS AMONG 16 VARIABLES FOR 166 CONTROL STUDENTS AT THE END OF THE FALL SEMESTER, 1963-64

Variable	ACT Eng	ACT Comp	SUI Read	CEEB Sept	COOP Eng Sept	Theme Rating September			%-ile Rank H.S.	CEEB Jan.	COOP Eng Jan.	Theme Rating January			GPA Fall Sem	Fall Comp Sem
						1	2	Total				1	2	Total		
ACT English																
ACT Composite	.67															
SUI Reading	.51	.65														
CEEB Sept.	.63	.64	.60													
COOP Eng. Sept.	.58	.46	.37	.60												
Sept. Theme:																
Reader 1	.39	.30	.26	.41	.38											
Reader 2	.34	.24	.23	.41	.34	.38										
Sum 1 & 2	.44	.33	.29	.49	.44	.87	.78									
%-ile Rank H.S.	.38	.34	.27	.37	.25	.14	.16	.18								
CEEB January	.69	.64	.55	.74	.60	.36	.40	.45	.36							
COOP Eng. Jan.	.65	.54	.45	.63	.72	.41	.41	.49	.38	.69						
Jan. Theme:																
Reader 1	.37	.29	.27	.30	.24	.26	.28	.32	.18	.38	.36					
Reader 2	.29	.29	.26	.28	.26	.24	.28	.30	.11	.32	.34	.39				
Sum 1 & 2	.40	.35	.32	.34	.30	.30	.33	.38	.18	.42	.42	.83	.83			
GPA Fall Sem.	.51	.60	.49	.45	.37	.27	.24	.31	.44	.42	.47	.22	.28	.30		
Fall Comp. Sem.	.54	.46	.37	.44	.40	.32	.21	.33	.33	.44	.55	.40	.31	.43	.64	

INTERCORRELATIONS AMONG 16 VARIABLES FOR 166 MATCHED PAIRS OF STUDENTS AT THE END OF THE FALL SEMESTER, 1963-64

Variables Involved

1 American College Testing Program (ACT) English standard score

2 American College Testing Program (ACT) Composite standard score

3 State University of Iowa (SUI) Reading Comprehension Form A raw score

4 College Entrance Examination Board English Composition Test Form KPL1 (September, 1963) standard rating

5 Cooperative English Tests: English Expression Form 1A converted score (September, 1963)

6 Theme rating assigned by Reader 1 (September, 1963)

7 Theme rating assigned by Reader 2 (September, 1963)

8 Sum of theme ratings assigned by Readers 1 and 2 (September, 1963)

9 Percentile rank in high school graduating class

10 College Entrance Examination Board English Composition Test Form KPL2 standard rating (January, 1964)

11 Cooperative English Tests: English Expression Form 1B converted score (January, 1964)

12 Theme rating assigned by Reader 1 (January, 1964)

13 Theme rating assigned by Reader 2 (January, 1964)

14 Sum of theme ratings assigned by Readers 1 and 2 (January, 1964)

15 ONLY FOR CONTROL STUDENTS Grade in first semester course in English composition, 1963-64

16 Over-all grade point average for fall semester 1963-64

TABLE XVII

CORRELATION BETWEEN SELECTED PAIRS OF VARIABLES FOR THE 113 MATCHED PAIRS OF STUDENTS COMPLETING THE SPRING SEMESTER, MAY, 1964

Correlation Between	r for N=226 (Exp. + Control)	r for N=113 (Experimental)	r for N=113 (Control)
May, 1964, Theme Total and September, 1963, ACT English	.28	.28	.28
September, 1963, COOP English	.17	.10	.25
September, 1963, CEEB English	.27	.24	.29
September, 1963, Theme Total	.24	.23	.24
May, 1964, COOP English and September, 1963, COOP English	.67	.72	.61
May, 1964, CEEB English and September, 1963, CEEB English	.64	.65	.47
Second Semester Freshman Composition Course Grade and ACT English			.13
COOP English			.10
CEEB English			.08
September, 1963, Theme Total			.17
Cumulative Grade Point Index in May, 1964, and ACT Composite	.44	.38	.51
High School Rank	.24	.30	.21

In May, 1964, the 113 control students completed their second semester of freshman composition. Their marks in this course showed a negligible relationship with September test scores. The highest r was .17 for September theme total.

May, 1965--End of Second Year

In Table XVIII the correlation facts are for the 31 matched pairs of students for whom full data were available through May, 1965 (end of sophomore year). It is realized that for such a small sample the obtained r's are relatively unreliable. For an N of 31, an r as large as .36 is required for significance at the .05 level (two-tailed interpretation). For an N of 62, the corresponding correlation value is .25.

One of the kinds of evidence contained in Table XVIII is the relationship between variables over a span of two academic years. The first section of the table shows correlations between end-of-sophomore year theme total rating and beginning-of-freshman-year indexes. The largest r was for September, 1963, ACT English: .45 for the combined subgroups (N=62).

The September, 1963-May, 1965, correlation for each of the two objective tests is presented in the second section. The COOP r was .74 for the combined subgroups; CEEB yielded a corresponding r of .57.

For the control subgroup there is evidence concerning end-of-sophomore-year theme total and of freshman year grade in the second semester English composition course. The obtained r of -.02 is interesting. In contrast to this low correlation between theme and course grade, the correlation between May, 1965, theme total and May, 1964, theme total was .47 (control subgroup).

RELIABILITY OF CRITERION MEASURES

Research in English composition hinges on the reliability and validity of the measuring instruments employed. This section presents certain evidence concerning the reliability of the three criterion tests employed in the present investigation.

Cooperative English Tests: English Expression

This instrument, published in 1960, is composed of two parts: "Part I: Effectiveness," thirty items; and "Part II:

TABLE XVIII

CORRELATION BETWEEN SELECTED PAIRS OF VARIABLES FOR THE 31 MATCHED PAIRS OF STUDENTS COMPLETING THE SPRING SEMESTER, MAY, 1965

Correlation Between	r for N=62 (Exp. + Control)	r for N=31 (Experimental)	r for N=31 (Control)
May, 1965, Theme Total and			
September, 1963, ACT English	.45	.44	.48
September, 1963, COOP English	.40	.41	.37
September, 1963, CEEB English	.30	.24	.37
September, 1963, Theme Total	.35	.38	.32
May, 1965, COOP English and			
September, 1963, COOP English	.74	.75	.72
May, 1965, CEEB English and			
September, 1963, CEEB English	.57	.54	.61
May, 1965, Theme Total and			
Grade in English Composition II Second Semester Freshman Year			-.02
May, 1964, Theme Total	.42	.39	.47

Mechanics," sixty items. The time limits are 15 minutes and 25 minutes respectively. A student's score is the total number of correct responses. This raw score is transformed into a Converted Score by means of a table provided by the publishers of the test. For Form 1A, the possible range in converted scores is from 115 (raw score of 0) to 191 (raw score of 90). For the two forms of the test (1A, 1B) recommended for use with college freshmen and sophomores, the investigators were able to find reliability facts only for the twelfth grade level. The correlation between parallel forms was .84 and the standard error of measurement was on the order of 4.00 converted score units.

The College Entrance Examination Board English Composition Test

This is one of the CEEB achievement tests. Evidence about the functioning of this instrument seems to be directly concerned with validity. This is reflected in one of the earlier reports on the instrument, which appeared with the title "Composition Test Shows High Validity on Reliable Criterion of Writing Ability" (2). The excellent 84-page report called The Measurement of Writing Ability (3) also dealt primarily with the validity of the College Entrance Examination Board English Composition Test (CEEB). It is realized that to achieve validity a test author must at the same time achieve reliability. A third source of information was The Sixth Mental Measurements Yearbook. Holland Roberts, one of the three reviewers of the test, commented on reliability: "For the composition test a Kuder-Richardson formula 20 reliability of .85 and a standard error of measurement of 39 is reported, indicating satisfactory discrimination among the members of the test group." (5:590)

The Theme

The theme test consisted of an impromptu paper 300-500 words in length. Students were allowed up to two hours to write the paper. A new topic was used at each testing session, but at no testing session was more than one topic provided. Typically, the topic consisted of a quotation set in a framework intended to link the topic and the student's experience (See Appendix A). Experimental and control students wrote the paper at the same time under similar conditions--usually in a period between 3:00 and 5:00 p.m.

Each theme was evaluated by two independent readers (see discussion p. 18). Each reader assigned each paper a numerical value on a nine-point scale. It is thus possible to examine the extent of between-reader agreement in assigned ratings.

The investigators analyzed the scores assigned to 1,070 students: essentially the total incoming freshman class for September, 1963, at the State College of Iowa. Table XIX is a frequency distribution of the amount of difference between the two ratings for each paper.

TABLE XIX

FREQUENCY DISTRIBUTION OF THE DIFFERENCE IN TWO INDEPENDENT RATINGS ASSIGNED TO EACH OF 1,070 THEMES

Difference in Two Theme Ratings	N
0	260
1	432
2	283
3	72
4	20
5	3

It will be noted from Table XIX that 260 of the 1,070 themes received the same rating by two readers working independently. Another 432 themes were rated within a point of one another. On only 23 of the 1,070 themes was there an inter-reader disparity of four points or more. As each rater was marking on a 9-point scale (9 high, 1 low), there was a potential disparity of 8 points (9 minus 1) between ratings. This analysis seems to suggest that the themes were evaluated with considerable consistency. In the light of this kind of analysis, the student scores--the total of the two independent ratings--appear to be sufficiently reliable.

The above straightforward analysis of the extent of agreement on independent ratings of the same theme is the most meaningful basis for thinking about the theme reliability. When one moves to the tricky problem of producing a reliability coefficient, interpretations are exceedingly complex. One way of obtaining a reliability estimate is to conceive of this as a single-test-form reliability situation, involving a 9-point test. We would actually be studying the consistency of two independent ratings of a single test. It is a "short test" in terms of maximum possible score. Table XIV, page 46, shows that for 910 students the Reader 1-Reader 2 r was .35. This is a conservative estimate of reader reliability.

Another way to express the reliability of the theme is to regard the sum of the two ratings as a total test score, and each rating as the score on a half test.* This would put it in the context of an 18-point test. Data on the difference between the two independent ratings of each of the themes may be used directly in arriving at a reliability estimate, using a procedure outlined by Rulon (6:99-103). The standard deviation of the distribution of differences in theme ratings was .970. This may be used as the estimated standard error of measurement. The standard deviation of the distribution of total theme scores for the entire freshman class was 2.64. The reliability coefficient is then computed from $r_{12} = 1 - \frac{.970^2}{2.64^2}$. This produces an r of .87, a spuriously high coefficient of reader reliability.

Different methods for estimating theme reliability lead to such varying coefficients (of reader reliability), and the relevance of each technique to the present situation is so difficult to assess, that the most meaningful analysis for present purposes seems to the investigators to be that presented in Table XIX in the form of a distribution of differences in ratings of themes by two independent readers.

It is of interest to note that readers, when giving a paper its second reading, showed a slight tendency to be more generous in their evaluation than in their first reading. On 283 papers, the independent evaluations differed by two points. In 126 instances the first reading was higher; in 157 instances the second reading was higher.

PERFORMANCE BY SEX AND ABILITY LEVEL

Performance by Sex

Since a systematic superiority of women on tests of composition ability is ordinarily expected, the proportions of males and females in the sample was determined by the proportions in the entire entering freshman class. Furthermore, sex was one of the factors used in establishing the matched pairs. To determine whether this presumed superiority of women was manifest in the present study, the investigators analyzed some of the data by sex. Table XX shows the facts for the 113 matched pairs of students for whom full data were available through the second semester of college (May, 1964). The student performance by sex is reported for the three testing points of the freshman year (September, 1963; January, 1964; and May, 1964) on the three criterion measures.

---

*For an interesting report of the use of multiple readings of themes see Godshalk, Swineford, and Coffman. (3)

On COOP, the mean performance of the 83 women in the experimental subgroup was consistently somewhat higher than that of the 30 men. This situation also prevailed for the control subgroup. The observed superiority seems to have been slightly greater in May (about three points) than in the previous September (about two points).

On CEEB the differences in mean performance by the male and female components of the sample are not consistently in the same direction. The males tested somewhat higher than the females in September, but by May the direction of superiority had been reversed. This is equivalent to saying that during the freshman year the observed mean gain by the female students was greater than that by the male students. Following is an analysis of these shifts. For the male subsamples (N=30), a t-value of 2.04 is required for significance at the .05 level, for the female subsamples (N=83), a t-value of 1.99. Thus the mean gain of the

Experimental

	N	Sept. 1963 CEEB		May 1964 CEEB		Difference May 1964 Minus Sept. 1963	r	t
		Mean	S.D.	Mean	S.D.			
Male	30	505.40	71.60	527.13	67.82	21.73	.54	1.78
Female	83	502.98	65.83	540.78	81.45	37.80	.70	5.85

Control

	N	Sept. 1963 CEEB		May 1964 CEEB		Difference May 1964 Minus Sept. 1963	r	t
		Mean	S.D.	Mean	S.D.			
Male	30	505.63	75.98	515.47	85.81	9.84	.73	.90
Female	83	495.47	69.65	527.67	78.40	32.20	.60	4.40

TABLE XX

PERFORMANCE OF 113 MATCHED PAIRS OF STUDENTS, BY SEX, ON THREE CRITERION MEASURES AT THE BEGINNING, MIDDLE, AND END OF FIRST YEAR OF COLLEGE

Subgroup and Sex	N	Testing	COOP English Test:English Expression (1960) Converted Score		CEEB English Composition Test Standard Rating		Theme: Sum of Two Ratings	
			Mean	S.D.	Mean	S.D.	Mean	S.D.
Experimental		September, 1963						
Men	30		159.83	8.26	505.40	71.60	8.10	2.51
Women	83		162.07	7.07	502.98	65.83	8.63	1.82
Total	113		161.48	7.47	503.62	67.42	8.49	2.04
Control		September, 1963						
Men	30		160.30	7.36	505.63	75.98	8.07	2.44
Women	83		162.55	6.54	495.47	69.65	8.64	1.82
Total	113		161.96	6.84	498.17	71.53	8.49	2.02
Experimental		January, 1964						
Men	30		164.10	7.57	477.97	73.83	8.31	2.28(N=29)
Women	83		165.48	6.84	509.59	80.96	9.58	2.35(N=81)
Total	113		165.12	7.06	501.19	80.35	9.25	2.40(N=110)
Control		January, 1964						
Men	30		164.37	6.20	519.97	70.14	9.43	2.43
Women	83		166.39	5.05	516.98	77.09	9.34	2.26(N=82)
Total	113		165.85	5.45	517.77	75.32	9.37	2.31(N=112)
Experimental		May, 1964						
Men	30		163.13	9.88	527.13	67.82	9.70	2.04
Women	83		167.14	7.03	540.78	81.45	9.82	2.04
Total	113		166.08	8.08	537.16	78.29	9.79	2.04
Control		May, 1964						
Men	30		163.63	7.97	515.47	85.81	9.10	2.50
Women	83		166.06	7.35	527.67	78.40	9.72	1.80
Total	113		165.42	7.60	524.46	80.61	9.56	2.02

females in both the experimental subgroup and the control subgroup was highly significant.

For the males in the experimental subgroup the mean gain was not quite significant, and in the control subgroup the mean gain did not approach significance. Clearly, women outgained men. For this subsample, the experimental men outgained control men on CEEB.

A superficial interpretation of this CEEB evidence is that women may be neither handicapped nor benefitted by instruction in freshman composition whereas men may be handicapped. There should be an investigation of the unusual possibility that the freshman composition course may have an inhibitory effect on the writing improvement of male students. Certainly these data suggest that in any composition study in which the ratio of males to females is not held constant for all treatment groups, "effects" may improperly be attributed to treatments instead of to the male-female imbalance among the groups.

On the theme ratings, reported in Table XX, there was a slight tendency for the females to score higher than the males. The one exception was in the control subgroup for the January, 1964, testing. In May, 1964, within the experimental subgroup, the mean for males was 9.70, and the mean for females was 9.82. Within the control subgroup, the means were 9.10 (male) and 9.72 (female).

The evidence summarized in this section indicates that on all three criterion measures the females did indeed tend to perform somewhat better than the males. On the CEEB test the comparisons of gain scores between September and May show an almost startling superiority for the females, especially in the control group. Had the females dominated one group and the males the other without the investigators' being aware of it, erroneous conclusions could easily have been drawn. The total evidence supports the wisdom of maintaining an equal ratio between the sexes in experimental and control groups in research concerning composition.

Performance by Ability Level

Another consideration in methods experiments is the possibility that the effect may not be the same at all levels of student ability. It is conceivable, for example, that in an investigation such as the present one, the omission of formal course work in composition might have a negative effect on low-ability students but not on high-ability students. The point involved is whether or not what is true over-all is true at specified ability levels.

When the research was planned, it was decided to include some analysis of data which would present the evidence for students at four levels of ability. To make this auxiliary analysis, the 113 matched pairs of students for whom complete data were available through May, 1964 (the freshman year) were used. It seemed desirable to establish the four ability levels by some measure which was at hand before the beginning of the freshman year. The ACT English scores were such measures. Records indicated that one could divide the total incoming freshman class in September, 1963, into four ability levels of essentially equal size by using the ACT standard score intervals of 25 and above, 23-24, 21-22, and 20 and below. To explore performance at different ability levels CEEB was used.

Table XXI presents the CEEB data for the 113 matched pairs grouped by the ACT standard score intervals. It may be noted, first of all, that the N's are not uniform. This variation in N's exists from level to level and, to a lesser extent, between experimentals and controls at each level. It would be anticipated that attrition would be more noticeable among the students in the lower-ability groups. This was true in the present situation except that the smallest N's were at the third rather than the fourth ability level. This pattern of frequencies was already present in the sample of 166 matched pairs which completed the first semester. In January, 26 of the 166 individuals (16%) in each of the two subgroups were located at the third ability level. Of the 113 matched pairs in May, 14 of the experimentals (12%) and 20 of the controls (18%) were at the third ability level.

Of principal interest in Table XXI is the column headed "CEEB May Minus Sept.," which contains information concerning the experimental-control relationship at each of four ability levels in addition to the facts concerning main effect. At how many, if any, of the four ability levels were the findings substantially different from the over-all findings? Only at the third level, the level with the smallest N's, is the evidence different from the over-all picture. For the 14 experimental students at the third ability level, the mean gain was 49.22, and for the 20 control s~~tudents~~, -8.20. This evidence may or may not be suggestive of actual interaction. At all four levels, and over-all, the experimentals are somewhat higher than the controls on September-May gain on the CEEB. The provocative fact is that at the third level this advantage is conspicuously greater than at the other

TABLE XXI

GAINS ON COLLEGE ENTRANCE EXAMINATION BOARD ENGLISH DURING COLLEGE FRESHMAN YEAR AT EACH OF FOUR ACT-ENGLISH ABILITY LEVELS

Subgroup	N	ACT Standard Score	ACT		CEEB Standard Rating September		CEEB Standard Rating May		CEEB May Minus Sept.	r (Sept. vs. May)
			Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	
Experimental	41	25+	26.17	1.08	545.49	69.94	581.12	71.34	35.63	.72
Control	37	25+	26.06	1.13	544.42	62.20	580.03	58.24	35.61	.37
Experimental	30	23,24	23.47	.50	499.87	38.17	542.30	61.91	42.43	.32
Control	32	23,24	23.59	.49	497.78	69.05	538.38	76.27	40.60	.69
Experimental	14	21,22	21.36	.48	484.71	68.35	533.93	77.14	49.22	.33
Control	20	21,22	21.70	.46	488.55	56.49	480.35	55.37	-8.20	.30
Experimental	28	20 or below	18.11	1.90	455.79	47.25	468.89	51.77	13.10	.43
Control	24	20 or below	18.92	1.38	439.76	48.29	461.92	66.52	22.16	.47
Experimental	113	Overall	22.86	3.37	503.62	67.42	537.16	78.29	33.54	.65
Control	113	Overall	23.01	2.83	498.17	71.53	524.46	80.61	26.29	.63

three levels. (See Table VII for overall CEEB data for 113 matched pairs on September, January, and May testing.)

If one looks at the evidence in Table XXI from the standpoint of student ability level, disregarding the treatment factor, it is evident that the amount of gain on CEEB, September to May, was at least as great in the upper one-half of the ability breakdown as it was in the lower one-half. This is in contrast to the view frequently held that the greater gains will be at the lower levels, the lower gains at the higher levels.

CONCLUSIONS AND OBSERVATIONS

The objective of the total research project was to test two hypotheses. The first of these was that the writing performance of students enrolled in a college freshman composition sequence is not significantly different from the writing performance of comparable students not enrolled in a college freshman composition sequence when the two subgroups have attended college for an equal length of time. The second hypothesis was that evidence from a single sample at one institution would be confirmed by evidence from a second sample at the given college and by evidence from other colleges and/or universities. The present discussion relates to the test of the first hypothesis. The test of the second hypothesis must await the analysis of evidence collected from the five institutions which replicated the study described in this report.

It should be emphasized that the investigators are presenting these conclusions and observations on the basis of the pilot phase of the study alone, without any utilization of evidence from the major phase--the phase involving replication by five institutions. This is the fair thing to do, and it is also the best way for the investigators to make maximum use of the pilot phase, one of the purposes of which was to provide experience useful in the major phase. It is clearly understood that the evidence for the major phase, to be reported in 1967, will constitute a far more dependable basis for conclusions and observations than does the evidence in the present document.

The Basic Findings

In the pilot phase, statistics were obtained for three evaluative measures applied at the beginning of the experiment and at three subsequent times. The hypothesis was thus subjected to review with each of three testing instruments on each of three testing occasions. The main subsamples (subgroups) were students, from an original 210 matched pairs, who remained in the experiment at least one semester; some of the students persisted through all four semesters involved in the pilot study. There were nine main "tests" of the null hypothesis. Only one of them led to a rejection of the hypothesis. The following tabular presentation shows these facts. Examination of the statistical portion of the present report and of this summary table forces the generalization that the first hypothesis has been sustained. The only point at which significant difference was found between major subgroups at a particular testing period was at the end of the first semester of instruction. The difference, in favor of the control students (those receiving instruction), is on CEEB, an objective test.

NINE MAIN COMPARISONS OF EXPERIMENTAL SUBGROUP AND CONTROL SUBGROUP:

Was the null hypothesis rejected?

	January, 1964 (N=166 Matched Pairs)	May, 1964 (N=113 Matched Pairs)	May, 1965 (N=31 Matched Pairs)
COOP ENG	No	No	No
CEEB ENG	Yes, Controls Excelled	No	No
THEME TOTAL	No	No	No

Among the 166 matched pairs who were tested in January, 1964, were the 113 matched pairs whose performance is reported for May, 1964, and among the 113 matched pairs were the 31 matched pairs whose performance is reported for May, 1965. The investigators report in the following table the January, 1964, performance of the subsamples of 113 and 31 matched pairs, and the May, 1964, performance of the subsample of 31 matched pairs.

NINE SECONDARY COMPARISONS OF EXPERIMENTAL SUBSAMPLE AND CONTROL SUBSAMPLE:

Was the null hypothesis rejected?

	January, 1964		May, 1964
	(N=113 Matched Pairs)	(N=31 Matched Pairs)	(N=31 Matched Pairs)
COOP ENG	No	Yes, Controls Excelled	No
CEEB ENG	Yes, Controls Excelled	No	No
THEME TOTAL	No	No	No

This table reveals that in the nine secondary comparisons two instances of significant difference appeared. The subsample of 113 matched pairs which persisted through May, 1964, showed a statistically significant difference in favor of the control students on their January, 1964, performance on CEEB, the same objective test on which the 166 control students presented its sole instance of a statistically significant difference. The subsample of 31 matched pairs who persisted to the end of the second year of the investigation achieved a significant difference on the January, 1964, performance on COOP. One may summarize by saying that in eighteen comparisons of performance--nine for the main subgroups and nine for the subsamples--the null hypothesis was rejected three times.

Performance by Sex and by Ability Level

The investigators explored selected factors in addition to those involved in the basic comparisons. One of these was the variation of performance by sex. Females as groups consistently performed somewhat better than males on the criterion measures. As sex was used as one of the matching criteria in the present investigation, the ratio of males to females was the same in the experimental and the control subgroups. If these ratios had not been kept constant, an observed superiority for a subgroup could have been improperly attributed to the treatment rather than to the ratio between sexes. An important conclusion is, then, that in investigations concerning competence in composition, the ratio between sexes must be taken into account in the groups whose performance is being studied.

The second factor concerning which the investigators made a special analysis was that of gains by ability levels. It is often assumed that the greatest improvement will be shown by the lower-average group, as they presumably have not only the "capacity" to improve but "room" on the evaluative instruments in which to show their progress. Conversely, the students at the upper levels cannot go much higher, and may even decline--at least as a result of the phenomenon of regression. This general assumption did not hold in the present study. Rather, the gains at the upper ability levels were greater than at the lower ability levels when the sample was segmented on the basis of ACT English, a test students took before they entered college. It would appear that the disparity in performance between the "better" and the "worse" students in composition tends to widen somewhat during the first year of college.

Observations

The first observation is that the only testing occasion on which significant differences between subgroups were found was in January, 1964, at the end of the participants' first semester of college. In each instance, the difference was in favor of the control subgroup on an objective test. The subgroups of 166 matched pairs and 113 matched pairs showed this difference on CEEB, and the subsample of 31 on COOP. Without now speculating about the difference between these two instruments, one may observe that this is the only testing occasion at which the students receiving instruction in freshman composition showed a significant superiority over the students not receiving such instruction. By the end of the second semester, the experimental students were not statistically different from the control students. Such performance suggests that the advantage of one semester of instruction in composition apparently disappears by the end of the first year even when instruction continues during the second semester. Put another way, it suggests that instruction hastens a development which will occur eventually through maturation or some other influence.

Related to the above is an observation concerning the performance in May, 1964, and in May, 1965, of the 31 matched pairs who persisted. Both the experimental subgroup and the control subgroup showed a decline in performance during the second year of college on CEEB. For the experimental subgroup, the decline was statistically significant. If, during the sophomore year, there is a decline in writing ability as measured by CEEB, it appears more likely to occur among those who have not had instruction in college freshman composition than among those who have had such instruction. This second-year evidence suggests that perhaps students reach a plateau of performance at the end of the freshman year. These are straws in the wind, observations made on the basis of one objective test--CEEB. Nevertheless, they are interesting hints, if nothing more.

SUMMARY

This is an interim report of Research Project 2188 as amended, which investigates the effectiveness of college freshman composition. The objective of the project is to test two hypotheses: (1) that the writing performance of students completing freshman composition does not differ significantly from the writing of students not taking freshman composition when both have been in college the same length of time, and (2) that replication of the experiment at several institutions will support the first hypothesis. This interim report covers the pilot phase of the investigation, developed at the State College of Iowa 1963-65, and relates to the first hypothesis only. Evidence of the test of the second hypothesis will be presented in the final report.

For the investigation, some students taking composition were matched with students not taking composition on the basis of age, sex, theme score, CEEB and COOP. The two objective tests (COOP and CEEB) and the theme were the criterion measures. Students were tested at the start (210 pairs), at the end of the first semester (166 pairs), at the end of the second semester (113 pairs), and at the end of the fourth semester (31 pairs). The themes were evaluated by teams selected by Fred Godshalk, Chairman of Test Development in Humanities at the Educational Testing Service, from the pool of theme readers used by ETS for college entrance and advanced placement.

Results sustained the first hypothesis, that the writing performance of a group of students who complete a year of composition (control) does not differ significantly from that of students who have had no composition (experimental). Of nine main comparisons--COOP, CEEB, and THEME on each of three occasions (end of first semester, end of second semester, end of fourth semester)--the null hypothesis was supported on eight, the one exception being that the control group excelled on the CEEB at the end of first semester; at the end of the second semester and at the end of the fourth semester, there was no significant difference.

Two other factors were examined: performance by sex and by ability level. Females performed consistently better than males on all criteria. From this study, it is inferred that the ratios between the sexes must be taken into account whenever composition competence is a factor in a reasearch project. Ability level segments were determined by scores on ACT English. On this basis, obtained gains at upper-ability levels were somewhat greater than those at lower-ability levels. In this study, it appears that disparity in performance between upper and lower ability students

tends to increase during the first year of college.

The only testing occasion when there was a significant difference between control and experimental subgroups was at the end of the first semester on an objective test. This difference favored the control subgroup. This advantage disappeared by the end of the second semester. Instruction seems to hasten a development in writing achievement which will occur anyway as the result of instruction in a college environment or some other factor. Based again on an objective test only (CEEB), there is a decline in performance during the second year, a decline which seems greater for those who have had no writing instruction than for those who have had such instruction.

REFERENCES

1. Braddock, Richard, Richard Lloyd-Jones, and Lowell Schoer. *Research in Written Composition*. Champaign, Illinois: *National Council of Teachers of English*, 1963.

2. "Composition Test Shows High Validity on Reliable Criterion of Writing Ability," *ETS Developments*, XI (January 1963), 1 & 4.

3. Godshalk, Fred, Frances Swineford, and William E. Coffman. *The Measurement of Writing Ability*. New York: College *Entrance Examination Board, 1966.*

4. Jewell, Ross M. and Gordon J. Rhum. *The Relative Effectiveness of Two Methods of Instruction in College Freshman Composition: Closed-Circuit Television and 'Normal' Classroom.* Cedar Falls, Iowa: State College of Iowa, February, 1966.

5. Roberts, Holland, [a review of the CEEB English Composition Test], *Sixth Mental Measurements Yearbook*. Ed. Oscar K. Buros. Highland Park, N. J.: Gryphon Press, 1965, pp. 589-91.

6. Rulon, Phillip J., "A Simplified Procedure for Determining the Reliability of a Test by Split-Halves," *Harvard Educational Review*, IX(1939), 99-103.

APPENDIX A

Theme Topics and Instructions

The principles followed in selecting topics, the use of a single topic on each test administration, and the equivalence of topics actually employed need to be discussed briefly.

Three criteria were established in selecting topics for the theme tests: the topic must be of a middle level of abstraction, it must be related to the students' experience, and it must call for an individual rather than a stock response. A middle level of abstraction avoided favoring either the students who were skillful in exploring general principles or the students who happened to have special knowledge related to a specific topic. A topic related to the students' experience and knowledge allowed them to support and illustrate their general statements with particulars readily available to them. A topic calling for an individual rather than a stock response provided a test of the students' ability to establish and support an original thesis.

The use of a single topic rather than a choice among several topics on each testing occasion avoided the introduction of an additional variable whose influence would be difficult to estimate. Such a restriction seemed justified by the fact that the students' performance as individuals was not under investigation. There is no reason to believe that if the students had had a choice of topics, comparison of their group performance would have been different from that resulting from a single topic.

Equivalence of topics across testing occasions was not vital, as students' change scores on theme performance were not considered in the conclusions in this study. Though it was hoped that the topics used would be comparable to one another, any lack of similarity which may be present cannot be used meaningfully in speculation about the results achieved. The subgroups were compared with one another on their performance at each testing occasion. Changes from occasion to occasion within subgroups were not investigated.

On the following pages are the instructions and theme topics for the various testing sessions. The complete instruction sheets, with places for the readers' ratings, the name and number of the student, and the like have not been reproduced as these details are irrelevant and their reproduction difficult. It should be noted, however, that the original instruction sheets were so arranged that the graders could learn neither the student's name nor the date on which the paper was written, and the second reader could not see the rating given the paper by the first reader.

(Theme Instructions for September 1963)

THEME INSTRUCTIONS

1. The paper which you are about to write will be judged on your success in presenting your thoughts in a clear, unified, well-organized manner, observing the conventions of standard written English. You should think about the topic until you have determined the idea you want to convey to the reader and the general procedure you will follow in doing so. Then you may write your paper. Do not hesitate to make a brief outline if you desire to do so (use the back of this sheet). An outline is not required.

2. You should be as neat as you can, but you should not hesitate to make changes if you believe them to be necessary. You do not have to make a fair copy.

3. WRITE ON ONE SIDE OF THE PAPER ONLY. If you need more paper, ask for it.

4. You may write in pen or in pencil, but pen is preferred.

5. Be certain that your NAME IS ON EVERY SHEET, in the UPPER RIGHT-HAND CORNER, and that it, as well as the rest of your writing, is as legible as you can make it.

6. Turn in all of the paper given to you.

7. LENGTH: 300-500 words.

TOPIC

Few question the idea that loyalty is a virtue. However, there are occasions when loyalties conflict. For example, loyalty to one's family or school may conflict with loyalty to one's friends; loyalty to an ideal may conflict with loyalty to the group. Thus one must sometimes choose to be disloyal to one thing in order to be loyal to another.

Attempt to determine a principle which you feel would be useful in making such a choice, using examples from your own experience and observation to indicate how you arrived at the principle you recommend.

(Theme Instructions for January 1964)

THEME INSTRUCTIONS

1. The paper which you are about to write will be judged on your success in presenting your thoughts in a clear, unified, well-organized manner, observing the conventions of standard written English. You should think about the topic until you have determined what idea you want to convey to the reader and the general procedure you will follow in doing so. Then you may write your paper. Do not hesitate to make a brief outline if you desire to do so (use the back of this sheet). An outline is not required.

2. You should write as neatly and legibly as you can, but you should not hesitate to make changes between the lines if you believe them to be necessary. You do not have to copy the paper over.

3. WRITE ON ONE SIDE OF THE PAPER ONLY. If you need more paper, ask for it.

4. You must write with INK OR BALL-POINT PEN.

5. Be certain to write your STUDENT NUMBER in each of the blanks (two at the top, one at the bottom) provided for it on this sheet, and in the upper right-hand corner of each page of your theme.

6. Turn in all of the paper given to you.

7. LENGTH: 300-500 words.

TOPIC

Conventional is a word frequently used to refer to customary attitudes, beliefs or actions. In the United States it is a convention for men to be clean-shaven, women to wear a certain amount of make-up, boys to be interested in sports and girls to be interested in becoming wives and mothers. A person who is unconventional in some way departs from the conventions of action or belief of the society of which he is a part.

With this explanation in mind, discuss the following statement: "Convention is society's safeguard, but also its potential executioner." To what extent and in what ways do you agree with this statement? Use examples and details from your knowledge and experience to support your conclusion.

(Theme Instructions for May 1964)

THEME INSTRUCTIONS

1. The paper which you are about to write will be judged on your success in presenting your thoughts in a clear, unified, well-organized manner, observing the conventions of standard written English. You should think about the topic until you have determined what idea you want to convey to the reader and the general procedure you will follow in doing so. Then you may write your paper. Do not hesitate to make a brief outline if you desire to do so (use the back of this sheet). An outline is not required.

2. You should write as neatly and legibly as you can, but you should not hesitate to make changes between the lines if you believe them to be necessary. You do not have to copy the paper over.

3. WRITE ON ONE SIDE OF THE PAPER ONLY. If you need more paper, ask for it.

4. You must write with INK OR BALL-POINT PEN.

5. Be certain to write your STUDENT NUMBER in each of the blanks (two at the top, one at the bottom) provided for it on this sheet, and in the upper right-hand corner of each page of your theme.

6. Turn in all of the paper given to you.

7. LENGTH: 300-500 words.

TOPIC

I returned and saw under the sun, that the race is not to the swift, nor the battle to the strong, neither yet bread to the wise, nor yet riches to men of understanding, nor yet favor to men of skill; but time and chance happeneth to them all.
--Ecclesiasties 9:11.

Using your experience and observation, indicate why you agree or disagree with the statement made in this quotation.

(Theme Instructions for May 1965)

THEME INSTRUCTIONS

1. The paper which you are about to write will be judged on your success in presenting your thoughts in a clear, unified, well-organized manner, observing the conventions of standard written English. You should think about the topic until you have determined what idea you want to convey to the reader and the general procedure you will follow in doing so. Then you may write your paper. Do not hesitate to make a brief outline if you desire to do so (use the back of this sheet). An outline is not required.

2. You should write as neatly and legibly as you can, but you should not hesitate to make changes between the lines if you believe them to be necessary. You do not have to copy the paper over.

3. WRITE ON ONE SIDE OF THE PAPER ONLY. If you need more paper, ask for it.

4. Begin on the third line of the first sheet, and WRITE ON EVERY LINE THEREAFTER.

5. You must write with INK or BALL-POINT PEN.

6. Be certain to write your STUDENT NUMBER in the blank provided at the top of this instruction sheet in the upper left-hand corner under the Total Score box. It should also be written on each page of your theme. Do NOT write your name, or the name of your school, in any place other than the blank provided at the bottom of this sheet.

7. Turn in all of the paper given to you.

8. You must stay at least one hour and fifteen minutes.

9. LENGTH: 300-500 words.

TOPIC

As society becomes increasingly complex, the number of people upon whom we are dependent increases. Daniel Boone killed a bear and ate it. When we buy steak, we purchase the services of the person who produced the animal, the person who fattened it, the person who took it to market, the packing company which bought it, slaughtered it, and dressed it, the trucker who transported it to the store from which we bought it, and, of course, the grocer himself. Each person must do his part if we are to have the steak. Even this picture is greatly over-simplified. There are, for example, the gasoline which fueled the truck and the truck itself. Considering the interdependence illustrated by the story of the steak, how free are we to guide our own lives? Are we liberated from stalking, killing, skinning, and cleaning our dinner, or are we robbed of our independence? Can we say, as Henley did, "I am the master of my fate; /I am the captain of my soul"? Does modern technology liberate us or dominate us? Present your opinion, based upon your knowledge, observation, and experience.

APPENDIX B

Choice of Experimental Design

In planning research, the most complex questions are those concerned with the choice of experimental design. The questions are both theoretical and functional. These two kinds of consideration come together when one finally must decide the best way, under the circumstances in which a given study will be made, to collect and analyze data for meaningful samples of students. In the present study, three circumstances dictated the choice of a matched pairs design.

The first circumstance was the college administration's stipulation that students who were to receive the experimental treatment be informed of the fact prior to their registration. It seemed essential that such students, their parents, and the faculty advisors be given advance information about the purpose of the research and its impact on them. These experimental students would not receive instruction in freshman composition--a major departure from normal college experience. Given the faith of students in the importance of composition (4:48), to have denied them enrollment on registration day without prior warning could have induced anxiety and resentment, possibly producing a kind of "reverse" Hawthorne effect. Added to this would have been confusion in registration, irritation among advisors, and concern among parents.

Thus the investigators were compelled to select, in advance of September registration, the students who would receive the experimental treatment. As described on page 17, this procedure involved selecting a pool of students from those who, by July 1, 1963, had met admission requirements and expressed their intention to enroll in the State College of Iowa. There was, of course, no assurance that all of the selected pool would actually enroll. This pool, which was a random sample from the July list, would not be a random sample of the September freshman class. That is, some entering freshman students had no opportunity to be included, and some who were included in the July group did not enroll.

A second circumstance was the duration of the investigation. The experimental design called for the students to be tested through the end of their sophomore year. That relatively heavy attrition would occur was certain;* that it would have an equal effect on both treatment groups seemed unlikely. Among other considerations, the control students would be enrolled in a course which frequently causes students trouble, while the experimental students would not. In any event, the possibility that attrition would occur in such a way that the two treatment groups would become progressively dissimilar could not be ignored.

Related to the attrition problem was the importance of maintaining the same ratio of males to females in both of the subgroups. The investigators believed, and their belief is supported by data subsequently examined (see page 56), that females would perform somewhat better than males on measures of composition ability. Should the ratio between sexes in one group become substantially different from the ratio in the other group, the likelihood of distorted results would be present.

A third circumstance was the audience which would read the research. As the investigation concerns the effectiveness of a course usually taught in departments of English, members of English departments would be the group for whom the report was primarily intended. It seems fair to say that such an audience would have considerable difficulty in following the intricacies of analysis of covariance. Though this consideration may at first seem somewhat frivolous, its pertinence to the potential impact of the project is nonetheless real.

In the light of these circumstances, the investigators became convinced that the matched-pairs design should be employed. Matching after September registration insured a list of students who were actually enrolled. Use of the matched pairs design with sex as one criterion made certain that the ratio between males and females would be the same for both subgroups not only at the beginning, but at any subsequent point in the investigation. Use of matched pairs minimized the possibility that in the attrition which would occur over the life of the experiment some factor would operate unequally to reduce the similarity of the subgroups. Finally, use of matched pairs enabled the investigators to present results in a manner which would make them readily available to members of English departments and directors of freshman composition.

---

*The Registrar of the State College of Iowa estimates that the attrition for the freshman class between September, 1963, and May, 1964, was on the order of 19 per cent, and the attrition between September, 1963, and May, 1965, was approximately 40 per cent.

The investigators could, of course, have set up the subgroups from among the students whose data were available in July, taking first a random sample of the total group, pairing them, and then for each matched pair of students randomly assigning one member of the pair to the experimental treatment and the other member of the pair to the control treatment. However, in July the only pertinent test data available for the students was their performance on ACT English. As the investigators wished to match as closely as possible, they decided to wait until more tests could be administered during the fall semester orientation period. Doing so permitted matching as reported on page 18, by age, sex, theme performance, and a score derived from performance on the CEEB and COCP. This precision in matching provided increased confidence in the similarity between the two treatment groups. Closeness in matching was also facilitated by the fact that the supply of subjects was greater in September than it was in July.

Three additional points. Since there were only two treatment groups, the matched pairs approach was more feasible than if there had been several treatment groups. Secondly, the investigators did not have to use, indeed did not wish to use, intact classroom groups for the control treatment. Finally, in methods experiments generally, random samples of a real population are not attainable. Near-randomness is achieved only in the beginning stages, and not in the groups which actually complete the experimental period.

APPENDIX C

Procedure for Evaluating Themes

Prior to each reading session, Mr. Jewell would send to Mr. Godshalk about forty themes, selected at random. From this sample Mr. Godshalk would determine the general nature of the total set of themes. He would choose a number of themes that in his judgment were typical of range and treatment, and Mr. Jewell would have these duplicated. These became the sample themes used during the reading as practice themes.

Mr. Godshalk's main responsibility when the raters (the smallest number was nine) had assembled was to communicate to them the criteria for evaluating the papers. First, he would have Mr. Cowley and Mr. Jewell describe the purpose of the investigation, the circumstances under which the papers had been written, and the students who had written them. He would then explain the rating scale. When all questions concerning its application had been answered, he would distribute several sample themes to be rated. After he had made a tally of the various values assigned to these papers, he would allow individuals to explain their ratings or to question his rating. If a rater seemed to be over-reacting to something in the papers, something which Mr. Godshalk believed from examination of the sample papers was typical, he would so inform the readers and caution them against misinterpreting particular aspects of the papers. For example, on the loyalty topic, he felt it useful to point out that Iowa has many strongly religious communities, so that the use of religious principles in response to the theme topic was the reflection of sincere belief and not pious platitude to impress the reader.

Before setting the readers to work in earnest, he would remind them that since they were experienced readers their first judgment of a theme as a whole was probably as valid as any subsequent judgment they might make of the same paper. Therefore, they were not to pause and consider but were to read and respond. As the rating session progressed, Mr. Godshalk would note whether any particular rater seemed to judge consistently in a way different from the other raters. At relatively frequent intervals, he would interrupt the reading to allow the readers to relax and would read aloud papers which had been passed on to him by individual readers. Frequently, these papers posed special problems which Mr. Godshalk would have the group discuss, always making clear his own judgment. The goal of the initial orientation and of the subsequent breaks in the reading was for Mr. Godshalk to convey to the readers his criteria and to get them to standardize their scoring so that they would agree in their ratings. The reading would be most "perfect" when all of the readers rated all the papers in the same way that Mr. Godshalk would rate them. In practice his standards would be slightly altered if a consensus indicated they should be. Thus, the validity of the evaluation could be no greater than the validity of Mr. Godshalk's criteria as modified on occasion by discussion with the readers.